Noname manuscript No.
(will be inserted by the editor)

Service Rate, Busy Period & Throughput Analysis of a Horizontal Traffic Queue

Mohammad Motie · Ketan Savla

Abstract We consider a horizontal traffic queue (HTQ) on a periodic road segment, where vehicles arrive according to a spatio-temporal Poisson process, and depart after traveling a distance that is sampled independently and identically from a spatial distribution. When inside the queue, the speed of a vehicle is proportional to a power $m > 0$ of the distance to the vehicle in front. The service rate of HTQ is equal to the sum of the speeds of the vehicles, and has a complex dependency on the state (vehicle locations) of the system. We show that the service-rate increases (resp., decreases) in between arrivals and departures for $m < 1$ (resp., $m > 1$) case. For a given initial condition, we define the throughput of such a queue as the largest arrival rate under which the queue length remains bounded. We extend the busy period calculations for M/G/1 queue to our setting, including for non-empty initial condition. These calculations are used to prove that the throughput for $m = 1$ case is equal to the inverse of the time required to travel average total distance by a solitary vehicle in the system, and also to derive a probabilistic upper bound on the queue length over a finite time horizon for the $m > 1$ case. Finally, we study throughput under a release control policy, where the additional expected waiting time caused by the control policy is interpreted as the magnitude of the perturbation to the arrival process. We derive a lower bound on throughput for a given combination of maximum allowable perturbation, for $m < 1$ and $m > 1$ cases. In particular, if the allowable perturbation is sufficiently large, then this lower bound grows unbounded as $m \to 0^+$. Illustrative simulation results are also presented.

1 Introduction

We consider a horizontal traffic queue (HTQ) on a periodic road segment, where vehicles arrive according to a spatio-temporal Poisson process, and depart the queue after traveling a distance that is sampled independently and identically from a spatial distribution. When inside the queue, the speed of a vehicle is proportional to a power $m > 0$ of the distance to the vehicle in front. For a given initial condition, we define the throughput of such a queue as the largest arrival rate under which the queue length remains bounded. We provide rigorous analysis for the service rate, busy period distribution, and throughput of the proposed HTQ.

Our motivation for studying HTQ comes from advancements in connected and autonomous vehicle technologies that allow to program individual vehicles with rules that can optimize system level

M. Motie
Sonny Astani Department of Civil and Environmental Engineering
University of Southern California
E-mail: motiesha@usc.edu

K. Savla
Sonny Astani Department of Civil and Environmental Engineering
University of Southern California
E-mail: ksavla@usc.edu

performance. Within this application context, one can interpret the results of this paper as rigorously characterizing the impact of a parametric class of car-following behavior on system throughput.

In the linear case ($m = 1$), i.e., when the speed of every vehicle is proportional to the distance to the vehicle directly in front, the periodicity of the road segment implies that the sum of the speeds of the vehicles is proportional to the total length of the road segment, i.e., it is constant. This feature allows us to exploit the equivalence between workload and queue length to show that, independent of the initial condition and almost surely, the throughput is the inverse of the time required by a solitary vehicle to travel average distance.

In the non-linear case ($m \neq 1$), the cumulative service rate of HTQ queue is constant if and only if all the inter-vehicle distances are equal. For all other inter-vehicle configurations, we show that the service rate is strictly decreasing (resp., strictly increasing) in the super-linear, i.e., $m > 1$ (resp., sub-linear, i.e., $m < 1$) case. The service rate exhibits an another contrasting behavior in the sub- and super-linear regimes. In the super-linear case, the service rate is maximum (resp., minimum) when all the vehicles are co-located (resp., when the inter-vehicle distances are equal), and vice-versa for the sub-linear case. Using a combination of these properties, we prove that, when the length of the road segment is at most one, the throughput in the super-linear (resp., sub-linear) case is upper (resp., lower) bounded by the throughput for the linear case.

We prove the remaining bounds on the throughput for the non-linear case as follows. The standard calculations for joint distributions of duration and number of arrivals during a busy period for M/G/1 queue are extended to the HTQ setting, including for non-empty initial conditions. These joint distributions are used to derive probabilistic upper bounds on queue length over finite time horizons for HTQ for the $m > 1$ case. Such bounds are optimized to get lower bounds on throughput defined over finite time horizons. Simulation results show good comparison between such lower bounds and numerical estimates.

We also analyze throughput in the sub-linear and super-linear cases under perturbation to the arrival process, which is attributed to the additional expected waiting time induced by a release control policy that adds appropriate delay to the arrival times to ensure a desired minimum inter-vehicle distance $\triangle > 0$ at the time of a vehicle joining the HTQ. Since the minimum inter-vehicle distance is non-decreasing in between arrivals and jumps, this implies an upper bound on the queue length which is inversely proportional to $\triangle$. We derive a lower bound on throughput for a given combination of maximum allowable perturbation. In particular, if the allowable perturbation is sufficiently large, then this lower bound grows unbounded, as $m \to 0^+$.

Queueing models have been used to model and analyze traffic systems. The focus here has been primarily on vertical queues, under which vehicles travel at maximum speed until they hit a congestion spot where all vehicles queue on top of each other. The queue length and waiting time of a minor traffic stream at an unsignalized intersection where major traffic stream has high priority is studied in [20] and [8]. In [9], a vertical single server queue is utilized to model the queue length distribution at signalized intersections. In [11], a state-dependent queuing system is used to model vehicular traffic flow where the service rate depends on the number of vehicles on each road link.

On the other hand, the *horizontal traffic queue* terminology has been primarily used to study macroscopic traffic flow, e.g., see [10]. While such models capture the macroscopic relationship between traffic flow and density, a rigorous description and analysis of an underlying queue model is lacking. Indeed, to the best of our knowledge, there is no prior work on the analysis of a traffic queue model that explicitly incorporates car-following behavior.

The proposed HTQ has an interesting connection with processor sharing (PS) queues, and this connection does not seem to have been documented before. A characteristic feature of PS queues is that all the outstanding jobs receive service simultaneously, while keeping the total service rate of the server constant. The simplest model is where the service rate for an individual job is equal to $1/N$, where N is the number of outstanding jobs. In our proposed system, one can interpret the road segment as a server simultaneously providing service to all the vehicles, with the service rate of an individual vehicle equal to its speed. This natural analogy between HTQ and PS queues, to the best of our knowledge,

was reported for the first time in our recent work [15]. The $1/N$ rule applied to our setting implies that all the vehicles travel with the same speed. Clearly, such a rule, or even the general discriminatory PS disciplines, e.g., see [14], are not applicable to the car following models considered in this paper. Indeed, the proposed HTQ is best described as a state-dependent PS queue.

In the PS queue literature, the focus has been on the sojourn time and queue length distribution. For example, see [17] and [21] for M/G/1-PS queue and [6] for G/G/1-PS queue. Fluid limit analysis for PS queue is provided in [4] and [7]. However, relatively less attention has been paid to the throughput analysis of state-dependent PS queues. In [16,12,5], throughput analysis for state-dependent PS queues is provided, where throughput is defined as the quantity of work achieved by the server per unit of time. Stability analysis for a single server queue with workload-dependent service and arrival rate is provided in [2] and [3]. However, the dependence of service rate on the system state in the HTQ proposed in the current paper is complex, and hence none of these results are readily applicable.

In summary, there are several novel contributions of the paper. First, we propose a novel horizontal traffic queue and place it in the context of processor-sharing queues and state-dependent queues. We establish monotonicity properties of service rates in between jumps (i.e., arrivals and departures), and derive bounds on change in service rates at jumps. Second, we adapt busy period calculations for M/G/1 queue to our current setup, including for non-empty initial conditions. These results allow us to provide tight results for throughput in the linear case, and probabilistic bounds on queue length over finite time horizon in the super-linear case. We also study throughput under a batch release control policy, whose effect is interpreted as a perturbation to the arrival process. We provide lower bound on the throughput for a maximum permissible perturbation for sub- and super-linear cases. In particular, we show that, for sufficiently large perturbation, this lower bound grows unbounded as $m \to 0^+$. It is interesting to compare our analytical results with simulation results, which suggest a sharp transition in the throughput from being unbounded in the sub-linear regime to being bounded in the super-linear regime. While our analytical results do not exhibit such a phase transition yet, their novelty is in providing rigorous estimates of any kind on the throughput of horizontal traffic queues under nonlinear car following models.

The rest of the paper is organized as follows. We conclude this section with key notations to be used throughout the paper. The setting for the proposed horizontal traffic queue and formal definition of throughput are provided in Section 2. Section 3 contains useful properties on the dynamics in service rate in between and during jumps. Key busy period properties for the M/G/1 queue are extended to the HTQ case in Section 4.2. Throughput analysis is reported in Section 5. Simulations are presented in 6. Concluding remarks and directions for future work are presented in Section 7. A few technical intermediate results are collected in the appendix.

Notations

Let $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ denote the set of real, non-negative real, and positive real numbers, respectively. Let $\mathbb{N}$ be the set of natural numbers. If x_1 and x_2 are of the same size, then $x_1 \geq x_2$ implies element-wise inequality between x_1 and x_2. If x_1 and x_2 are of different sizes, then $x_1 \geq x_2$ implies inequality only between elements which are common to x_1 and x_2 – such a common set of elements will be specified explicitly. For a set $\mathcal{J}$, let $\mathrm{int}(\mathcal{J})$ and $|\mathcal{J}|$ denote the interior and cardinality of $\mathcal{J}$, respectively. Given $a \in \mathbb{R}$, and $b > 0$, we let $\mod(a, b) := a - \lfloor \frac{a}{b} \rfloor b$. Let $\mathcal{S}_N^L$ be the $N-1$-simplex over L, i.e., $\mathcal{S}_N^L = \left\{x \in \mathbb{R}_+^N \mid \sum_{i=1}^N x_i = L\right\}$. When $L = 1$, we shall use the shorthand notation $\mathcal{S}_N$. When referring to the set $\{1, \ldots, N\}$, for brevity, we let the indices $i = -1$ and $i = N+1$ correspond to $i = N$ and $i = 1$ respectively. Also, for $p, q \in \mathcal{S}_N$, we let $D(p||q)$ denote the K-L divergence of q from p, i.e., $D(p||q) := \sum_{i=1}^N p_i \log(p_i/q_i)$. We also define a permutation matrix, $P^- \in \{0,1\}^{N \times N}$, as follows:

$$P^- := \begin{bmatrix} \mathbf{0}_{N-1}^T & 1 \\ I_{N-1} & \mathbf{0}_{N-1} \end{bmatrix}$$

where $\mathbf{0}_N$ and $\mathbf{1}_N$ stand for vectors of size N, all of whose entries are zero and one, respectively. We shall drop N from $\mathbf{0}_N$ and $\mathbf{1}_N$ whenever it is clear from the context.

2 The Horizontal Traffic Queue (HTQ) Setup

Consider a periodic road segment of length L; without loss of generality, we assume it be a circle. Starting from an arbitrary point on the circle, we assign coordinates in $[0, L]$ to the circle in the clockwise direction (See Figure 1). Vehicles arrive on the circle according to a spatio-temporal process: the arrival process $\{A(t), t \geq 0\}$, is assumed to be a Poisson process with rate $\lambda > 0$, and the arrival locations are sampled independently and identically from a spatial distribution φ and mean value $\bar{\varphi}$. Without loss of generality, let the support of φ be $\mathrm{supp}(\varphi) = [0, \ell]$ for some $\ell \in [0, L]$. Upon arriving, vehicle i travels distance d_i in a counter-clockwise direction, after which it departs the system. The travel distances $\{d_i\}_{i=1}^{\infty}$ are sampled independently and identically from a spatial distribution ψ with support $[0, R]$ and mean value $\bar{\psi}$. Let the set of φ and ψ satisfying the above conditions be denoted by Φ and Ψ respectively. The stochastic processes for arrival times, arrival locations, and travel distances are all assumed to be independent of each other.

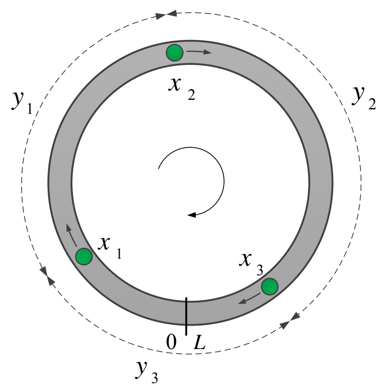


Fig. 1 Illustration of the proposed HTQ with three vehicles.

2.1 Dynamics of vehicle coordinates between jumps

Let the time epochs corresponding to arrival and departure of vehicles be denoted as $\{\tau_1, \tau_2, \ldots\}$. We shall refer to these events succinctly as *jumps*. We now formally state the dynamics under this car-following model. We describe the dynamics over an arbitrary time interval of the kind $[\tau_j, \tau_{j+1})$. Let $N \in \mathbb{N}$ be the fixed number of vehicles in the system during this time interval. Define the inter-vehicle distances associated with vehicle coordinates $x \in [0, L]^N$ as follows:

$$y_i(x) = \mod(x_{i+1} - x_i, L), \quad i \in \{1, \ldots, N\} \tag{1}$$

where we implicitly let $x_{N+1} \equiv x_1$ (See Figure 1 for an illustration). Note that the normalized inter-vehicle distances y/L are probability vectors. When inside the queue, the speed of every vehicle is proportional to a power $m > 0$ of the distance to the vehicle directly in front of it. We assume that this power $m > 0$ is the same for every vehicle at all times. Then, starting with $x(\tau_j) \in [0, L]^N$, the vehicle coordinates over $[\tau_j, \tau_{j+1})$ are given by:

$$x_i(t) = \mod\left(x_i(\tau_j) + \int_{\tau_j}^{t} y_i^m(x(z))\, dz, L\right), \qquad \forall\, i \in \{1, \ldots, N\}, \quad \forall\, t \in [\tau_j, \tau_{j+1}), \tag{2}$$

Remark 1 It is easy to see that the clock-wise ordering of the vehicles is invariant under (1)-(2).

The dynamics in inter-vehicle distances is given by:

$$\dot{y}_i = y_{i+1}^m - y_i^m, \qquad i \in \{1, \ldots, N\} \tag{3}$$

where we implicitly let $y_{N+1} \equiv y_1$.

2.2 Change in vehicle coordinates during jumps

Let $x(\tau_j^-) = \big(x_1(\tau_j^-), \ldots, x_N(\tau_j^-)\big) \in [0, L]^N$ be the vehicle coordinates just before the jump at τ_j. If the jump corresponds to the departure of vehicle $k \in \{1, \ldots, N\}$, then the coordinates of the vehicles $x(\tau_j) = (x_1(\tau_j), \ldots, x_{N-1}(\tau_j)) \in [0, L]^{N-1}$ after re-ordering due to the jump, for $i \in \{1, \ldots, N-1\}$, are given by:

$$x_i(\tau_j) = \begin{cases} x_i(\tau_j^-) & i \in \{1, \ldots, k-1\} \\ \\ x_{i+1}(\tau_j^-) & i \in \{k+1, \ldots, N-1\}\,. \end{cases}$$

Analogously, if the jump corresponds to arrival of a vehicle at location $z \in [0, \ell]$ in between the locations of the k-th and $k+1$-th vehicles at time τ_j^-, then the coordinates of the vehicles $x(\tau_j) = (x_1(\tau_j), \ldots, x_{N+1}(\tau_j)) \in [0, L]^{N+1}$ after re-ordering due to the jump, for $i \in \{1, \ldots, N+1\}$, are given by:

$$\begin{aligned} x_{k+1}(\tau_j) &= z \\ x_i(\tau_j) &= \begin{cases} x_i(\tau_j^-) & i \in \{1, \ldots, k\} \\ \\ x_{i-1}(\tau_j^-) & i \in \{k+2, \ldots, N+1\}\,. \end{cases} \end{aligned}$$

2.3 Problem statement

Let $x_0 \in [0, L]^{n_0}$ be the initial coordinates of n_0 vehicles present at $t = 0$. An HTQ is described by the tuple $(L, m, \lambda, \varphi, \psi, x_0)$. Let $N(t; L, m, \lambda, \varphi, \psi, x_0)$ be the corresponding queue length, i.e., the number of vehicles at time t for an HTQ $(L, m, \lambda, \varphi, \psi, x_0)$. For brevity in notation, at times, we shall not show the dependence of N on parameters which are clear from the context.

In this paper, our objective is to provide rigorous characterizations of the dynamics of the proposed HTQ. A key quantity that we study is throughput, defined below.

Definition 1 (Throughput of HTQ) Given $L > 0$, $m > 0$, $\varphi \in \Phi, \psi \in \Psi$, $x_0 \in [0, L]^{n_0}$, $n_0 \in \mathbb{N}$ and $\delta \in [0, 1)$, the throughput of HTQ is defined as:

$$\lambda_{\max}(L, m, \varphi, \psi, x_0, \delta) := \sup \left\{\lambda \geq 0 : \; \Pr\left(N(t; L, m, \lambda, \varphi, \psi, x_0) < +\infty, \quad \forall t \geq 0\right) \geq 1 - \delta\right\}. \tag{4}$$

Figure 2 shows the complex dependency of throughput on key queue parameters such as m and L. In particular, it shows that for every L, φ, ψ, x_0 and φ, the throughput exhibits a phase transition from being unbounded for $m \in (0, 1)$ to being bounded for $m > 1$. Moreover, Figure 2 also suggests that, for sufficiently small L, throughput is monotonically non-increasing in m, and that it is monotonically non-decreasing in $m > 1$, for sufficiently large L. Also, it can be observed that initial condition can also affect the throughput. We now develop analytical results that match the throughput profile in Figure 2 as closely as possible. To that purpose, we will make extensive use of novel properties of *service rate* and *busy period* of the proposed HTQ, which could be of independent interest.

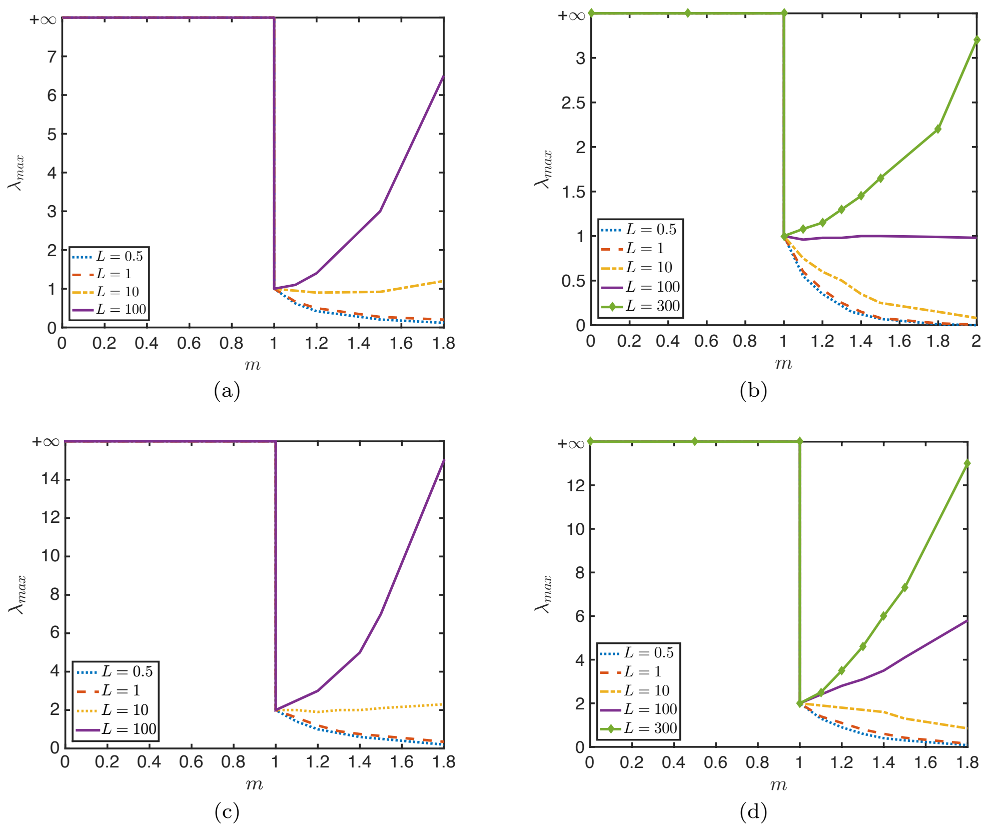


Fig. 2 Throughput for various combinations of m, L, and n_0. The parameters used in individual cases are: (a) $\varphi = \delta_0$, $\psi = \delta_L$, and $n_0 = 0$ (b) $\varphi = \delta_0$, $\psi = \delta_L$, and $n_0 = 100$ (c) $\varphi = U_{[0,L]}$, $\psi = U_{[0,L]}$, and $n_0 = 0$ (d) $\varphi = U_{[0,L]}$, $\psi = U_{[0,L]}$, and $n_0 = 100$. In all the cases, the locations of initial n_0 vehicles were chosen at equal spacing in $[0, L]$.

3 Service Rate Properties of the Horizontal Traffic Queue

For every $y \in \mathcal{S}_N^L$, $N \in \mathbb{N}$, $L > 0$, we let $y_{\min} := \min_{i \in \{1,\dots,N\}} y_i$, and $y_{\max} := \max_{i \in \{1,\dots,N\}} y_i$ denote the minimum and maximum inter-vehicle distances respectively. It is easy to establish the following monotonicity properties of $y_{\min}$ and $y_{\max}$.

Lemma 1 (Inter-vehicle Distance Monotonicity Between Jumps) *For any $y \in \mathcal{S}_N^L$, $N \in \mathbb{N}$, $L > 0$, under the dynamics in* (3)*, for all $m > 0$*

$$\frac{d}{dt} y_{\min} \geq 0 \qquad \& \qquad \frac{d}{dt} y_{\max} \leq 0.$$

Proof Let $y_{\min}(t) = y_j(t)$, i.e., the j-th vehicle has the minimum inter-vehicle distance at time $t \geq 0$. Therefore, (3) implies that $\dot{y}_{\min}(t) = \dot{y}_j(t) = y_{j+1}^m(t) - y_j^m(t) \geq 0$. One can similarly show that $y_{\max}$ is non-increasing. $\square$

Due to the complex state-dependence of the departure process, the queue length process is difficult to analyze. We propose to study a related scalar quantity, called *workload* formally defined as follows, where we recall the notations introduced in Section 2.

Definition 2 (Workload) The workload associated with the HTQ at any instant is the sum of the distances remaining to be travelled by all the vehicles present at that instant. That is, if the current

coordinates and departure coordinates of all vehicles are $x \in [0, L]^N$ and $q \in \mathbb{R}_+^N$ respectively, with $q \geq x$, then the workload is given by:

$$w(x, q) := \sum_{i=1}^{N} (q_i - x_i).$$

Since the maximum distance to be travelled by any vehicle from the time of arrival to the time of departure is upper bounded by R, we have the following simple relationship between workload and queue length at any time instant:

$$w(t) \leq N(t)\, R\,, \qquad \forall\, t \geq 0\,. \tag{5}$$

An implication of (5) is that unbounded workload implies unbounded queue length in our setting. We shall use this relationship to establish an upper bound on the throughput. However, a finite workload does not necessarily imply finite queue length. In order to see this, consider the state of the queue with N vehicles, all of whom have distance $1/N$ remaining to be travelled. Therefore, the workload at this instant is $1/N \times N = 1$, which is independent of N.

When the workload is positive, its rate of decrease is equal to *service rate* in between jumps, defined next.

Definition 3 (Service Rate) When the HTQ is not idle, its instantaneous service rate is equal to the sum of the speeds of the vehicles present in the system at that time instant, i.e., $s(x) = \sum_{i=1}^{N} y_i^m(x)$.

Since the service rate depends only on the inter-vehicle distances, we shall alternately denote it as $s(y)$. For $m = 1$, $s(y) = \sum_{i=1}^{N} y_i \equiv L$, i.e., the service rate is independent of the state of the system, and is constant in between and during jumps. This property does not hold true in the nonlinear ($m \neq 1$) case. Nevertheless, one can prove interesting properties for the service rate dynamics. We start by deriving bounds on service rate in between jumps.

Lemma 2 (Bounds on Service Rates) *For any $y \in \mathcal{S}_N^L$, $N \in \mathbb{N}$, $L > 0$, under the dynamics in* (3)*,*

1. $L^m N^{1-m} \leq s(y) \leq L^m$ *if $m > 1$;*
2. $L^m \leq s(y) \leq L^m N^{1-m}$ *if $m \in (0, 1)$.*

Proof Normalizing the inter-vehicular distances by L, the service rate can be rewritten as

$$s(y) = L^m \sum_{i=1}^{N} \left(\frac{y_i}{L}\right)^m. \tag{6}$$

Therefore, for $m > 1$, $s(y) \leq L^m \sum_{i=1}^{N} \frac{y_i}{L} = L^m$. One can similarly show that, for $m \in (0, 1)$, $s(y) \geq L^m$. In order to prove the remaining bounds, we note that $\sum_{i=1}^{N} z_i^m$ is strictly convex in $z = [z_1, \ldots, z_N]$ for $m > 1$, and that the minimum of $\sum_{i=1}^{N} z_i^m$ over $z \in \mathcal{S}_N$ occurs at $z = \mathbf{1}/N$, and is equal to N^{1-m}. Similarly, for $m \in (0, 1)$, $\sum_{i=1}^{N} z_i^m$ is strictly concave in z, and its maximum over $z \in \mathcal{S}_N$ occurs at $z = \mathbf{1}/N$, and is equal to N^{1-m}. Combining these facts with (6), and noting that $y/L \in \mathcal{S}_N$, gives the lemma. □

Lemma 3 (Service Rate Monotonicity Between Jumps) *For any $y \in \mathcal{S}_N^L$, $N \in \mathbb{N}$, $L > 0$, under the dynamics in* (3)*,*

$$\frac{d}{dt} s(y) \leq 0 \quad \text{if } m > 1 \qquad \& \qquad \frac{d}{dt} s(y) \geq 0 \quad \text{if } m \in (0, 1)\,,$$

where the equality holds true if and only if $y = \frac{L}{N}\mathbf{1}$.

Proof The time derivative of service rate is given by:

$$\begin{aligned}\frac{d}{dt}s(y) &= \frac{d}{dt}\sum_{i=1}^{N} y_i^m = m\sum_{i=1}^{N} y_i^{m-1}\dot{y}_i \\ &= m\sum_{i=1}^{N} y_i^{m-1}\left(y_{i+1}^m - y_i^m\right) \end{aligned} \tag{7}$$

where the second equality follows by (3). The result then follows by application of Lemma 12, and by noting that $g(z) = z^m$ is a strictly increasing function for all $m > 0$, and $h(z) = z^{m-1}$ is strictly decreasing if $m \in (0,1)$, and strictly increasing if $m > 1$. □

The following lemma quantifies the change in service rate due to departure of a vehicle.

Lemma 4 (Change in Service Rate at Departures) *Consider the departure of a vehicle that changes inter-vehicle distances from $y \in \mathcal{S}_N^L$ to $y^- \in \mathcal{S}_{N-1}^L$, for some $N \in \mathbb{N} \setminus \{1\}$, $L > 0$. If $y_1 \geq 0$ and $y_2 \geq 0$ denote the inter-vehicle distances behind and in front of the departing vehicle respectively, at the moment of departure, then the change in service rate due to the departure satisfies the following bounds:*

1. *if $m > 1$, then $0 \leq s(y^-) - s(y) \leq (y_1 + y_2)^m \left(1 - 2^{1-m}\right)$;*
2. *if $m \in (0,1)$, then $0 \leq s(y) - s(y^-) \leq \min\{y_1^m, y_2^m\}$.*

Proof If $m > 1$, then $\left(\frac{y_1}{y_1+y_2}\right)^m + \left(\frac{y_2}{y_1+y_2}\right)^m \leq \frac{y_1}{y_1+y_2} + \frac{y_2}{y_1+y_2} = 1$, i.e., $s(y^-) - s(y) = (y_1 + y_2)^m - y_1^m - y_2^m \geq 0$. One can similarly show that $s(y) - s(y^-) \geq 0$ if $m \in (0,1)$.

In order to show the upper bound on $s(y^-) - s(y)$ for $m > 1$, we note that the minimum value of $z^m + (1-z)^m$ over $z \in [0,1]$ for $m > 1$ is 2^{1-m}, and it occurs at $z = 1/2$. Therefore,

$$\begin{aligned} s(y^-) - s(y) = (y_1 + y_2)^m - y_1^m - y_2^m &= (y_1 + y_2)^m \left(1 - \left(\frac{y_1}{y_1+y_2}\right)^m - \left(\frac{y_2}{y_1+y_2}\right)^m\right) \\ &\leq (y_1 + y_2)^m \left(1 - 2^{1-m}\right) \end{aligned}$$

The upper bound on $s(y) - s(y^-)$ for $m \in (0,1)$ can be proven as follows. Since $y_1^m \leq (y_1 + y_2)^m$, $s(y) - s(y^-) = y_1^m + y_2^m - (y_1 + y_2)^m \leq y_2^m$. Similarly, $s(y) - s(y^-) \leq y_1^m$. Combining, we get $s(y) - s(y^-) \leq \min\{y_1^m, y_2^m\}$. Note that, in proving this, we nowhere used the fact that $m \in (0,1)$. However, this bound is useful only for $m \in (0,1)$. □

Remark 2 (Change in Service Rate at Arrivals) The bounds derived in Lemma 4 can be trivially used to prove the following bounds for change in service rate at arrivals:

1. if $m > 1$, then $0 \leq s(y) - s(y^+) \leq (y_1 + y_2)^m \left(1 - 2^{1-m}\right)$;
2. if $m \in (0,1)$, then $0 \leq s(y^+) - s(y) \leq \min\{y_1^m, y_2^m\}$,

where y_1 and y_2 are the inter-vehicle distances behind and in front of the arriving vehicle respectively, at the moment of arrival.

The following lemma will facilitate generalization of Lemma 3. In preparation for the lemma, let $f(y,m) := m\sum_{i=1}^{N} y_i^{m-1}\left(y_{i+1}^m - y_i^m\right)$ be the time derivative of service rate, as given in (7).

Lemma 5 *For all $y \in int(\mathcal{S}_N^L)$, $N \in \mathbb{N} \setminus \{1\}$, $L > 0$:*

$$\frac{\partial}{\partial m} f(y,m)|_{m=1} = -LD\left(\frac{y}{L}||P^-\frac{y}{L}\right) \leq 0 \tag{8}$$

Additionally, if $L < e^{-2}$, then

$$\frac{\partial^2}{\partial m^2} f(y,m)|_{m=1} \geq 0 \tag{9}$$

Moreover, equality holds true in (8) and (9) if and only if $y = \frac{L}{N}\mathbf{1}$.

Proof Taking the partial derivative of $f(y,m)$ with respect to m, we get that

$$\frac{\partial}{\partial m} f(y,m) = \frac{f(y,m)}{m} + m \sum_{i=1}^{N} \left(y_i^{m-1} y_{i+1}^{m} \left(\log y_i + \log y_{i+1}\right) - 2 y_i^{2m-1} \log y_i \right)$$

In particular, for $m = 1$:

$$\begin{aligned} \frac{\partial}{\partial m} f(y,m)|_{m=1} &= f(y,1) + \sum_{i=1}^{N} \left(y_{i+1} \left(\log y_i + \log y_{i+1}\right) - 2 y_i \log y_i \right) \\ &= L \sum_{i=1}^{N} \frac{y_i}{L} \log \left(\frac{y_{i-1}/L}{y_i/L} \right) \\ &= -LD \left(\frac{y}{L} || P^{-} \frac{y}{L} \right) \end{aligned}$$

where, for the second equality, we used the trivial fact that $f(y,1) = 0$. Taking second partial derivative of $f(y,m)$ w.r.t. m gives:

$$\begin{aligned} \frac{\partial^2}{\partial m^2} f(y,m) = &\sum_{i=1}^{N} y_i^{m-1} \log y_i \left(y_{i+1}^{m} - y_i^{m} \right) + \sum_{i=1}^{N} y_i^{m-1} \left(y_{i+1}^{m} \log y_{i+1} - y_i^{m} \log y_i \right) \\ &+ \sum_{i=1}^{N} \left(y_i^{m-1} y_{i+1}^{m} \left(\log y_i + \log y_{i+1}\right) - 2 y_i^{2m-1} \log y_i \right) \\ &+ m \sum_{i=1}^{N} \left(y_i^{m-1} y_{i+1}^{m} \left(\log y_i + \log y_{i+1}\right)^2 - 4 y_i^{2m-1} \log^2 y_i \right) \end{aligned}$$

In particular, for $m = 1$:

$$\begin{aligned} \frac{\partial^2}{\partial m^2} f(y,m)|_{m=1} = &\sum_{i=1}^{N} \left(y_{i+1} - y_i \right) \log y_i + \sum_{i=1}^{N} \left(y_{i+1} \log y_{i+1} - y_i \log y_i \right) \\ &+ \sum_{i=1}^{N} \left(y_{i+1} \left(\log y_i + \log y_{i+1}\right) - 2 y_i \log y_i \right) \\ &+ \sum_{i=1}^{N} \left(y_{i+1} (\log y_i + \log y_{i+1})^2 - 4 y_i \log^2 y_i \right) \\ = &\sum_{i=1}^{N} \log^2 y_i \left(y_{i+1} - y_i \right) + 2 \sum_{i=1}^{N} \log y_i \left(y_{i+1} \log y_{i+1} + y_{i+1} - y_i \log y_i - y_i \right) \qquad (10) \\ \geq &\, 0 \end{aligned}$$

It is easy to check that, $\log z$, $\log^2 z$ and $z + z \log z$ are strictly increasing, strictly decreasing and strictly decreasing functions, respectively, for $z \in (0, e^{-2})$. Therefore, Lemma 12 implies that each of the two terms in (10) is non-negative, and hence the lemma. □

Lemma 5 implies that, for sufficiently small L, $f(y,m)$ is locally convex in m. One can use this property along with an exact expression for $\frac{\partial}{\partial m} f(y,m)$ in Lemma 5 at $m = 1$, and the fact that $f(y,1) = 0$ for all y, to develop a linear approximation in m of $f(y,m)$ around $m = 1$. The following lemma derives this approximation, as also suggested by Figure 3.

Lemma 6 *For a given $y \in int(\mathcal{S}_N^L)$, $n \in \mathbb{N}$, $L \in (0, e^{-2})$, there exists $\underline{m}(y) \in [0,1)$ such that*

$$\frac{d}{dt} s(y) \geq 2 \frac{(1-m)}{L} \left(y_{\max} - y_{\min} \right)^2, \qquad \forall\, m \in [\underline{m}(y), 1]$$

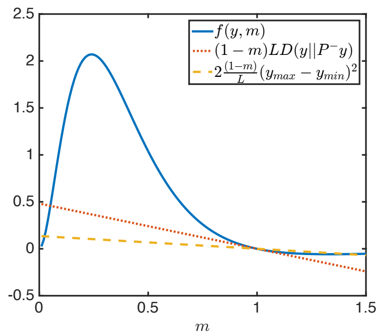


Fig. 3 $f(y,m)$ vs. m for a typical $y \in \mathcal{S}_{10}$.

Proof For a given $y \in \mathrm{int}(\mathcal{S}_N^L)$, the local convexity of $f(y,m) := \frac{d}{dt}s(y)$ in m, and the expression of $\frac{\partial}{\partial m}f(y,m)$ at $m=1$ in Lemma 5 implies that $\frac{d}{dt}s(y) \geq (1-m)LD\left(\frac{y}{L}||P^{-}\frac{y}{L}\right)$ for sufficiently small $m<1$. Pinsker's inequality implies $D\left(\frac{y}{L}||P^{-}\frac{y}{L}\right) \geq \frac{\|y-P^{-}y\|_1^2}{2L^2}$. This, combined with the fact that $\|y-P^{-}y\|_1 \geq 2(y_{\max}-y_{\min})$ for all $y \in \mathrm{int}(\mathcal{S}_N^L)$, gives the lemma. □

4 Busy Period Properties of the Horizontal Traffic Queue

The system is called *busy* when there is at least one vehicle on the road, or equivalently, the workload is positive. Once the system gets empty, it becomes *idle* up to the time of next arrival. Thus, the system alternates between busy and idle periods. Accordingly, while the first busy period might start from a non-zero initial condition, if the first busy period terminates, then the subsequent busy periods will start from the zero initial condition. In this paper, unless otherwise stated explicitly, we shall implicitly assume a zero initial condition when referring to a busy period.

4.1 Expected Busy Period Duration

The next lemma provides an expression for the expectation of the busy period duration in the linear case.

Lemma 7 *For any $\lambda < L/\bar{\psi}$, $L>0$, $m=1$, $\varphi \in \Phi$, $\psi \in \Psi$, the mean value of the busy period duration is equal to $\bar{\psi}/(L-\lambda\bar{\psi})$.*

Proof A busy period, say of duration B, is initiated by the arrival of a vehicle, say j, when the system is idle. Let the number of vehicles that arrive during the busy period be N_{bn}. Note that N_{bn} does not include the vehicle initiating the busy period. Therefore, the workload brought into the system during the busy period is equal to $w_B = \sum_{i=j}^{j+N_{bn}} d_i$. The expected value of N_{bn} can be obtained by conditioning on the duration of the busy period:

$$E[N_{bn}] = E\left[E[N_{bn}|B]\right] = E[\lambda B] = \lambda E[B] \tag{11}$$

where the second equality follows from the fact that the arrival process is a Poisson process. Since the event $\{N_{bn}+1=n\}$ is independent of $\{d_{j+i}, i>n\}$, $N_{bn}+1$ is a stopping time for the sequence $\{d_{j+i}, i \geq 1\}$. Therefore, using Wald's equation, e.g., see [18, Theorem 3.3.2], and (11), the expected value of the workload w_B added to the system during the busy period B is given by:

$$E[w_B] = \left(E[N_{bn}]+1\right)\bar{\psi} = \left(\lambda E[B]+1\right)\bar{\psi}. \tag{12}$$

Since the workload decreases at a constant rate L during a busy period, we have $B = w_B/L$ (see Figure 4 for an illustration). Therefore, $E[B] = E[w_B]/L$, which when combined with (12), establishes the lemma. □

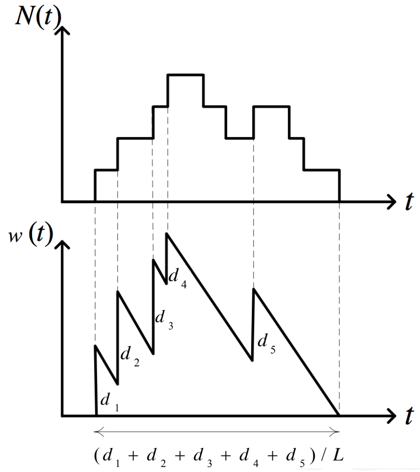


Fig. 4 (a) Queue length process and (b) workload process during a busy period.

Remark 3 Since the mean busy period duration is an upper bound on the mean waiting time, Lemma 7 also gives an upper bound on the mean waiting time. One can then use Little's law [13][1] to show that the mean queue length is upper bounded by $\lambda\bar{\psi}/(L - \lambda\bar{\psi})$.

Let $\mathcal{I}(t) := \int_0^t \delta_{\{w(s)=0\}} \mathrm{d}s$ be the cumulative *idle time* up to time t. The following result characterizes the long run proportion of the idle time in the linear case.

Proposition 1 *For any $\lambda < L/\bar{\psi}$, $m = 1$, $L > 0$, $\varphi \in \Phi, \psi \in \Psi$, the long-run proportion of time in which HTQ is idle is given by the following:*

$$\lim_{t\to\infty} \frac{\mathcal{I}(t)}{t} = 1 - \frac{\lambda\bar{\psi}}{L} > 0 \quad a.s.$$

Proof HTQ alternates between busy and idle periods. Let $Z = I + B$ be the duration of a cycle that contains an idle period of length I followed by a busy period of length B. Idle period, I, has the same distribution as inter-arrival times i.e. an exponential random variable with mean $1/\lambda$, and the mean value of B is given in Lemma 7. Note that duration of cycles, Z, are i.i.d. random variables. Thus, the busy-idle profile of the system is an alternating renewal process where renewals correspond to the moments at which the system gets idle. Suppose the system earns reward at a rate of one per unit of time when it is idle (and thus the reward for a cycle equals the idle time of that cycle i.e. I). Then, the total reward earned up to time t is equal to the total idle time in $[0, t]$ (or $\mathcal{I}(t)$), and by the result for renewal reward process (see [18], Theorem 3.6.1), with probability one, $\lim_{t\to\infty} \mathcal{I}(t)/t = E[I]/(E[B]+E[I])$. □

4.2 Busy Period Distribution

In this section, we compute the cumulative distribution function for the number of new arrivals during a busy period for a HTQ with constant service rate, say $p > 0$. This could, e.g., correspond to (3) for $m = 1$. However, our analysis in this section, is not restricted to this specific model, but applies to any HTQ with constant service rate p. This cumulative distribution for the number of new arrivals during a busy period, while of independent interest, will be used to derive lower bounds on the throughput in the super-linear case in Section 5.3. Our analysis is inspired by that of M/G/1 queue, e.g., see [18], where our consideration for non-zero initial condition appears to be novel.

[1] Little's law has previously been used in the context of processor sharing queues, e.g., in [1].

Let us consider an arbitrary busy period spanning time interval $(0, t)$, without loss of generality. For non-zero initial condition, one has to distinguish between the first and subsequent busy periods. Let the workload at the beginning of the arbitrary busy period, denoted as d_0, be sampled from θ. The relationship between θ and ψ is as follows. If the system starts with a non-zero initial initial condition with initial workload $w_0 > 0$, then the value of the d_0 for the first busy period will be deterministic and equals w_0, and hence $\theta = \delta_{w_0}$. However, for subsequent busy periods, or if the initial condition is zero, d_0 is sampled from $\theta = \psi$. The workload brought to the system by arriving vehicles, $\{d_i\}_{i=1}^{\infty}$, equals to the distance that vehicles wish to travel and are sampled identically and independently from the distribution ψ. When the system is busy, the workload decreases at a given constant rate $p > 0$. The busy period ends when the workload becomes zero.

Remark 4 We emphasize that d_0 denotes the workload at the beginning of a busy period (see Figure 5 for further illustration), and hence is not equal to zero when the queue starts from a zero initial condition.

In order to align our calculations with the standard M/G/1 framework, where service rate is assumed to be unity, we consider normalized workloads, $\tilde{d}_i := d_i/p$ for all $i \in \{0, 1, \cdots\}$ (see Figure 5 for an illustration). Correspondingly, let the distributions for the normalized distances be denoted as $\tilde{\theta}$ and $\tilde{\psi}$. Let the arrival time of the k-th new vehicle during $(0, t)$ be denoted as T_k, and let N_{bn} denote the number of arrivals in $(0, t)$, i.e., the total number of arrivals over the entire duration of the busy period, including the vehicle which initiates the busy period, is $N_{bn} + 1$.

A busy period ends at time t, and $N_{bn} = n - 1$ if and only if,

(i) $T_k \leq \tilde{d}_0 + \cdots + \tilde{d}_{k-1}, \qquad k = 1, \cdots, n-1$
(ii) $\tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t$
(iii) There are exactly $n - 1$ arrivals in $(0, t)$

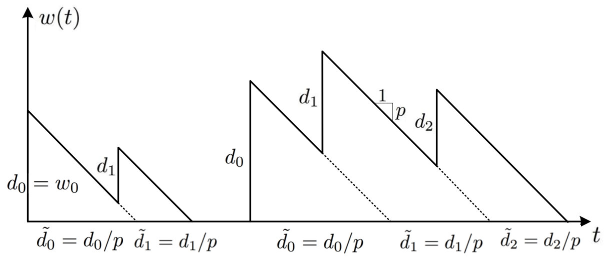


Fig. 5 Evolution of workload during first two busy periods for an HTQ with constant service rate p, and starting from a non-zero initial condition. In the first busy period, d_0 is equal to the workload w_0 associated with the non-zero initial condition. In the second busy period, d_0 is equal to the workload brought by the first vehicle that initiates that busy period.

By treating densities as if they are probabilities, we get:

$$
\begin{aligned}
&\Pr(B = t \text{ and } N_{bn} = n-1) \\
&= \Pr(\tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, n-1 \text{ arrivals in } (0,t), T_k \leq \tilde{d}_0 + \cdots + \tilde{d}_{k-1}, k = 1, \cdots, n-1) \\
&= \int_0^t \Pr(T_k \leq \tilde{d}_0 + \cdots + \tilde{d}_{k-1}, k = 1, \cdots, n-1 | n-1 \text{ arrivals in } (0,t), \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z) \\
&\quad \times \Pr(n-1 \text{ arrivals in } (0,t), \tilde{d}_1 + \cdots + \tilde{d}_{n-1} = t - z)\, \tilde{\theta}(z)\, \mathrm{d}z
\end{aligned} \tag{13}
$$

where we recall that B is the random variable corresponding to the busy period duration. By the independence of normalized distances and the arrival process, the second probability term in the integrand

in (13) can be expressed as

$$\Pr(n-1 \text{ arrivals in } (0,t), \tilde{d}_1 + \cdots + \tilde{d}_{n-1} = t-z) = e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}\tilde{\psi}_{n-1}(t-z) \tag{14}$$

where $\tilde{\psi}_n$ is the n-fold convolution of $\tilde{\psi}$ with itself.

In the first probability term in (13), it is given that the system receives $n-1$ arrivals in $(0,t)$ and since the arrival process is a Poisson process, the ordered arrival times, $\{T_1, T_2, \cdots, T_{n-1}\}$, are distributed as the ordered values of a set of $n-1$ independent uniform $(0,t)$ random variables $\{a_1, a_2, \cdots, a_{n-1}\}$ (see Theorem 2.3.1 in [18]). Thus,

$$\begin{aligned}&\Pr(T_k \le \tilde{d}_0 + \cdots + \tilde{d}_{k-1}, k = 1, \cdots, n-1 | n-1 \text{ arrivals in } (0,t), \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z)\\ &\quad = \Pr(a_k \le \tilde{d}_0 + \cdots + \tilde{d}_{k-1}, k = 1, \cdots, n-1 | \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z)\end{aligned} \tag{15}$$

By noting that $t-U$ will also be a uniform $(0,t)$ random variable whenever U is, it follows that $a_1, \cdots, a_{n-1}$ has the same joint distribution as $t-a_{n-1}, \cdots, t-a_1$. Thus, replacing a_k with a_{n-k} for $k \in \{1, \cdots, n-1\}$ in (15), we get

$$\begin{aligned}&\Pr(a_k \le \tilde{d}_0 + \cdots + \tilde{d}_{k-1}, k = 1, \cdots, n-1 | \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z)\\ &\quad = \Pr(t - a_{n-k} \le \tilde{d}_0 + \cdots + \tilde{d}_{k-1}, k = 1, \cdots, n-1 | \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z)\\ &\quad = \Pr(t - a_{n-k} \le t - (\tilde{d}_k + \cdots + \tilde{d}_{n-1}), k = 1, \cdots, n-1 | \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z)\\ &\quad = \Pr(a_{n-k} \ge \tilde{d}_k + \cdots + \tilde{d}_{n-1}, k = 1, \cdots, n-1 | \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z)\\ &\quad = \Pr(a_{n-k} \ge \tilde{d}_k + \cdots + \tilde{d}_{n-1}, k = 1, \cdots, n-1 | \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z) = \begin{cases} z/t & z < t \\ 0 & \text{otherwise}\end{cases}\end{aligned} \tag{16}$$

where the last equality follows from Lemma 14. If we let $H_p(t,n,\theta) := \Pr\{B \le t, N_{bn} = n-1\}$ when the service rate equals p, and d_0 has distribution θ; then, by plugging (14) and (16) in (13), we get

$$\frac{d}{dt}H_p(t,n,\theta) = e^{-\lambda t}\frac{(\lambda t)^{n-1}}{t(n-1)!}\int_0^t z\tilde{\psi}_{n-1}(t-z)\tilde{\theta}(z)\mathrm{d}z$$

By recalling the two special cases of interest to us: $\theta = \delta_{w_0}$ for a given non-zero initial workload w_0, and $\theta = \psi$ for zero initial condition, and using Lemma 13, we get that

$$G_p(t,n,\theta) := \frac{d}{dt}H_p(t,n,\theta) = \begin{cases} e^{-\lambda t}\frac{(\lambda t)^{n-1}w_0}{t(n-1)!p}\tilde{\psi}_{n-1}(t - w_0/p) & \theta = \delta_{w_0} \\ e^{-\lambda t}\frac{(\lambda t)^{n-1}}{n!}\tilde{\psi}_n(t) & \theta = \psi\end{cases} \tag{17}$$

For $r \in \mathbb{N}$, let $G_{r,p}(t,n,\theta)$ be the r-fold convolution of $G_p(t,n,\theta)$, defined in (17), with respect to t. In words, $G_{r,p}(t,n,\theta)$ is the probability that the number of new arrivals in each of (any) r busy periods is equal to $n-1$, and that the sum of durations of all the busy periods is equal to t. Similarly, for non-zero initial condition, let $G_{p_1}(\theta_1) * G_{r-1,p_2}(\theta_2)(t,n)$ be the probability that the number of new arrivals in each of (any) r busy periods is equal to $n-1$, and that the sum of durations of all the busy periods is equal to t, when the constant service rate for the first busy period is p_1 and is p_2 for the rest of the $r-1$ busy periods.

5 Throughput Analysis

5.1 Linear Case: $m = 1$

In this section, we provide an exact characterization of throughput for the linear case, i.e., when $m = 1$. Recall that, for $m = 1$, the service rate $s(y) = \sum_{i=1}^{N} y_i \equiv L$ is constant.

Proposition 2 *For any $L > 0$, $\varphi \in \Phi$, $\psi \in \Psi$, $x_0 \in [0, L]^{n_0}$, $n_0 \in \mathbb{N}$ and :*

$$\lambda_{max}(L, m = 1, \varphi, \psi, x_0, \delta = 0) \leq L/\bar{\psi}\,.$$

Proof By contradiction, assume $\lambda_{\max} > L/\bar{\psi}$. Let $r(t) := \sum_{i=1}^{A(t)} d_i$ be the workload added to the system by the $A(t)$ vehicles that arrive over $[0, t]$. Therefore,

$$w(t) = w_0 + r(t) - L(t - \mathcal{I}(t)) \tag{18}$$

where w_0 is the initial workload. The process $\{r(t),\, t \geq 0\}$ is a renewal reward process, where the renewals correspond to arrivals of vehicles and the rewards correspond to the distances $\{d_i\}_{i=1}^{\infty}$ that vehicles wish to travel in the system upon arrival before their departures. Inter-arrival times are exponential random variables with mean $1/\lambda$, and the reward associated with each renewal is independently and identically sampled from ψ, whose mean is $\bar{\psi}$. Therefore, e.g., [18, Theorem 3.6.1] implies that, with probability one,

$$\lim_{t \to \infty} \frac{r(t)}{t} = \lambda\bar{\psi} \tag{19}$$

Thus, for all $\varepsilon \in \big(0, \lambda\bar{\psi} - L\big)$, there exists a $t_0 \geq 0$ such that, with probability one,

$$\frac{r(t)}{t} \geq \lambda\bar{\psi} - \varepsilon/2 > L + \varepsilon/2 \qquad \forall\, t \geq t_0. \tag{20}$$

Since w_0 and $\mathcal{I}(t)$ are both non-negative, (18) implies that $w(t) \geq r(t) - Lt$ for all $t \geq 0$. This combined with (20) implies that, with probability one, $w(t) \geq \varepsilon t/2$ for all $t \geq t_0$, and hence $\lim_{t\to\infty} w(t) = +\infty$. This combined with (5) implies that, with probability one, $\lim_{t\to\infty} N(t) = +\infty$. □

Theorem 1 *For any $L > 0$, $\varphi \in \Phi$, $\psi \in \Psi$, $x_0 \in [0, L]^{n_0}$, $n_0 \in \mathbb{N}$:*

$$\lambda_{max}(L, m = 1, \varphi, \psi, x_0, \delta = 1) = L/\bar{\psi}\,.$$

Proof Assume that for some $\lambda < L/\bar{\psi}$, there exists some initial condition (x_0, n_0) such that the queue length grows unbounded with some positive probability. Since the workload brought by every vehicle is i.i.d., and the inter-arrival times are exponential, without loss of generality, we can assume that the queue length never becomes zero. That is, the idle time satisfies $\mathcal{I}(t) \equiv 0$. Moreover, (19) implies that, for every $\varepsilon \in \big(0, L - \lambda\bar{\psi}\big)$, there exists $t_0 \geq 0$ such that, with probability one,

$$\frac{r(t)}{t} \leq \lambda\bar{\psi} + \varepsilon/2 < L - \varepsilon/2 \qquad \forall t \geq t_0 \tag{21}$$

Combining (18) with (21), and substituting $\mathcal{I}(t) \equiv 0$, we get $w(t) < w_0 - \varepsilon t/2$, which implies that workload, and hence queue length, goes to zero in finite time after t_0, leading to a contradiction. Combining this with the upper bound proven in Proposition 2 gives the result. □

Remark 5 Theorem 1 implies that the throughput in the linear case is equal to the inverse of the time required to travel average total distance by a solitary vehicle in the system. In the linear case, the throughput can be characterized with probability one, independent of the initial condition of the queue.

5.2 Monotonicity of Throughput in m and x_0

In this section, we show the following monotonicity property of $\lambda_{\max}$ with respect to m for small values of L: for given $x_0 \in [0, L]^{n_0}$, $n_0 \in \mathbb{N}$, $L \in (0, 1)$, $\varphi \in \Phi$, and $\psi \in \Psi$, throughput is a monotonically decreasing function of m. For this section, we rewrite (2) in $\mathbb{R}_+^N$, i.e., without projecting onto $[0, L]^N$. Specifically, let the vehicle coordinates be given by the solution of

$$\dot{x}_i = y_i^m, \qquad x_i(0) = x_{0,i}, \qquad i \in \{1, \ldots, N\} \tag{22}$$

Let $X(t; x_0, m)$ denote the solution to (22) at t starting from x_0 at $t = 0$. We will compare $X(t; x_0, m)$ under different values of m and initial conditions x_0, over an interval of the kind $[0, \tau)$, in between arrivals and departures. We recall the notation that, if x_0^1 and x_0^2 are vectors of different sizes, then $x_0^1 \le x_0^2$ implies element-wise inequality only for components which are common to x_0^1 and x_0^2. In Lemma 8 and Proposition 3, this common set of components corresponds to the set of vehicles common between x_0^1 and x_0^2.

Lemma 8 *For any $L \in (0, 1]$, $x_0^1 \in \mathbb{R}_+^{n_1}$, $x_0^2 \in \mathbb{R}_+^{n_2}$, $n_1, n_2 \in \mathbb{N}$,*

$$x_0^1 \le x_0^2,\ n_2 \le n_1,\ 0 < m_2 \le m_1 \implies X(t; x_0^1, m_1) \le X(t; x_0^2, m_2) \qquad \forall\, t \in [0, \tau)$$

Proof The proof is straightforward when $n_1 = n_2$. This is because, in this case, since $y_i \le L \le 1$, $m_2 \le m_1$ implies $y_i^{m_2} \ge y_i^{m_1}$ for all $i \in \{1, \ldots, n_1\}$. Using this with Lemmas 10 and 11 gives the result.

In order to prove the result for $n_2 < n_1$, we show that $X(t; x_0^1, m_1) \le X(t; x_0^2, m_1) \le X(t; x_0^2, m_2)$. Note that the second inequality follows from the previous case. Therefore, it remains to prove the first inequality. Let $(i_1, \ldots, i_{n_2})$ be the set of indices of n_2 vehicles such that $0 \le x_{0,i_1}^2 \le \ldots \le x_{0,i_{n_2}}^2 \le L$. Similarly, let $(i_1, i_1{+}1, \ldots, i_2, i_2{+}1, \ldots)$ be the indices of n_1 vehicles in the order of increasing coordinates in x_0^1. Our assumption on the initial condition implies that $x_{0,i_k}^1 \le x_{0,i_k}^2$ for all $k \in \{1, \ldots, n_2\}$. For brevity, let $x^1(t) \equiv X(t; x_0^1, m_1)$, and $x^2(t) \equiv X(t; x_0^2, m_1)$. It is easy to check that, for all $t \in [0, \tau)$, and all $k \in \{1, \ldots, n_2\}$,

$$\dot{x}_{i_k}^1 = \left(x_{i_k+1}^1 - x_{i_k}^1\right)^{m_1} \le \left(x_{i_{k+1}}^1 - x_{i_k}^1\right)^{m_1} \tag{23}$$

Let $t \in [0, \tau)$ be the first time instant when $x_{i_k}^1(t) = x_{i_k}^2(t)$ for some $k \in \{1, \ldots, n_2\}$. Then, recalling $x_{i_{k+1}}^1(t) \le x_{i_{k+1}}^2(t)$, (23) implies that $\dot{x}_{i_k}^1(t) \le \left(x_{i_{k+1}}^2 - x_{i_k}^2\right)^{m_1} = \dot{x}_{i_k}^2(t)$. The result then follows from Lemma 10. □

Lemma 8 is used to establish monotonicity of throughput as follows.

Proposition 3 *For any $L \in (0, 1]$, $\varphi \in \Phi$, $\psi \in \Psi$, $\delta \in (0, 1)$, $x_0^1 \in [0, L]^{n_1}$, $x_0^2 \in [0, L]^{n_2}$, $n_1, n_2 \in \mathbb{N}$:*

$$x_0^1 \le x_0^2,\ n_2 \le n_1,\ 0 < m_2 \le m_1 \implies \lambda_{max}(L, m_1, \varphi, \psi, x_0^1, \delta) \le \lambda_{max}(L, m_2, \varphi, \psi, x_0^2, \delta)$$

Proof For brevity in notation, we refer to the queue corresponding to m_1, and initial condition x_0^1 as HTQ-S. We refer to the other queue as HTQ-F. Let λ, φ and ψ common to HTQ-S and HTQ-F be given. Let $x^1(t) \equiv X(t; x_0^1, m_1)$ and $x^2(t) \equiv X(t; x_0^2, m_2)$, and let $N_s(t)$ and $N_f(t)$ be the queue lengths in the two queues at time t. It suffices to show that $N_s(t) \ge N_f(t)$ for a given realization of arrival times, arrival locations, and travel distances. In particular, this also implies that the departure locations are also the same for every vehicle, including the vehicles present at $t = 0$, in both the queues.

Indeed, it is sufficient to show that $x^1(\tau) \le x^2(\tau)$ and $N_s(\tau) \ge N_f(\tau)$ where τ is the time of first arrival or departure from either HTQ-S or HTQ-F. Accordingly, we consider two cases, corresponding to whether τ corresponds to arrival or departure.

Since $x^1(t) \le x^2(t)$ for all $t \in [0, \tau)$ from Lemma 8, and the departure locations of all the vehicles in HTQ-S and HTQ-F are identical, the first departure from HTQ-S can not happen before the first departure in HTQ-F. Therefore, $N_s(\tau) \ge N_f(\tau)$. Since $x^1(\tau^-) \le x^2(\tau^-)$, and $x^2(\tau)$ is a subset of $x^2(\tau^-)$, we also have $x^1(\tau) \le x^2(\tau)$.

When τ corresponds to the time of the first arrival, since the arrivals happen at the same location in HTQ-S and HTQ-F, and since $x^1(\tau^-) \leq x^2(\tau^-)$, rearrangement of the indices of the vehicles to include the new arrival at $t = \tau$ implies that $x^1(\tau) \leq x^2(\tau)$. Moreover, since $N_s(\tau^-) \geq N_f(\tau^-)$, and the arrivals happen simultaneously in both HTQ-S and HTQ-F, we have $N_s(\tau) \leq N_f(\tau)$. □

Remark 6 Proposition 3 establishes monotonicity of throughput only for $L \in (0, 1]$. This is consistent with our simulation studies, e.g., as reported in Figure 2, according to which, the throughput is non-monotonic for large L.

For the analysis of the linear car following model, we exploited the fact that the total service rate of the system is constant. However, for the nonlinear model, i.e., $m \neq 1$, the total service rate depends on the number and relative locations of vehicles. The state dependent service rate of nonlinear models makes the throughput analysis much more complex. In the next section, we find probabilistic bound on the throughput in the super-linear case.

5.3 Throughput Bounds for the Super-linear Case from Busy Period Calculations

In this section, we derive lower bound on the throughput for the super-linear case. The next result computes a bound on the probability that the queue length of the HTQ satisfies a given upper bound over a given time interval, using the probability distribution functions from (17). In Propositions 4 and 5, for the sake of clarity, we add explicit dependence on λ to this probability distribution function.

Proposition 4 *For any $m > 1$, $M \in \mathbb{N}$, $L > 0$, $\lambda > 0$, $\varphi \in \Phi$, $\psi \in \Psi$, and zero initial condition $x_0 = 0$, the probability that the queue length is upper bounded by M over a given time interval $[0, T]$ satisfies the following bound:*

$$\Pr\big(N(t) \leq M \quad \forall t \in [0, T]\big) \geq \sup_{r \in \mathbb{N}} \sum_{n=1}^{M} \int_{T}^{\infty} G_{r, L^m M^{1-m}}(t, n, \psi, \lambda)\, \mathrm{d}t \tag{24}$$

Proof Let us denote the current queueing system as HTQ-f. We shall compare queue lengths between HTQ-f and a slower queueing system HTQ-s, which starts from the same (zero) initial condition, and experiences the same realizations of arrival times, locations and travel distances. Let every incoming vehicle into HTQ-s and HTQ-f be tagged with a unique identifier. At time t, let $\mathcal{J}(t)$ be the set of identifiers of vehicles present both in HTQ-s and HTQ-f, $\mathcal{J}_{s/f}(t)$ be the set of identifiers of vehicles present only in HTQ-s, and $\mathcal{J}_{f/s}(t)$ be the set of identifiers of vehicles present only in HTQ-f. Let v_i^f denote the speed of the vehicle in HTQ-f with identifier $i \in \mathcal{J}(t) \cup \mathcal{J}_{f/s}(t)$, as determined by the car-following behavior underlying (2). The vehicle speeds in HTQ-s are not governed by the car following behavior, but are rather related to the speeds of vehicles in HTQ-f as:

$$v_i^s(t) = \begin{cases} v_i^f(t) \dfrac{p}{v^f(t)} \dfrac{|\mathcal{J}(t)|}{|\mathcal{J}(t)| + |\mathcal{J}_{s/f}(t)|} & i \in \mathcal{J}(t) \\[2ex] \dfrac{p}{|\mathcal{J}(t)| + |\mathcal{J}_{s/f}(t)|} & i \in \mathcal{J}_{s/f}(t) \end{cases} \tag{25}$$

where $v^f(t) := \sum_{i \in \mathcal{J}(t)} v_i^f(t)$ is the sum of speeds of vehicles in HTQ-f that are also present in HTQ-s at time t, and p is a parameter to be specified. Indeed, note that $\sum_{i \in \mathcal{J}(t) \cup \mathcal{J}_{s/f}(t)} v_i^s(t) \equiv p$, i.e., p is the (constant) service rate of HTQ-s.

Consider a realization where the number of arrivals into HTQ-s with $p = L^m M^{1-m}$ during any busy period overlapping with $[0, T]$ does not exceed M. We refer to such a realization as *event* in the rest of the proof. Since the maximum queue length during a busy period is trivially upper bounded by the number of arrivals during that busy period, conditioned on the event, we have

$$N_s(t) \leq M, \qquad t \in [0, T] \tag{26}$$

Consider the union of departure epochs from HTQ-s and HTQ-f in $[0, T]$: $0 = \tau_0 \leq \tau_1 \leq \ldots$. If $\mathcal{J}_{f/s}(\tau_k) = \emptyset$ for some $k \geq 0$, then $\mathcal{J}_{f/s}(t) = \emptyset$ for all $t \in (\tau_k, \tau_{k+1})$. Hence, the service rate for HTQ-f over the interval (τ_k, τ_{k+1}) is $v^f(t)$, which, conditioned on the event, is lower bounded by $L^m M^{1-m} = p$ by Lemma 2. Therefore, $p/v^f(t) \leq 1$ over (τ_k, τ_{k+1}), and hence (25) implies that all the vehicles with identifiers in $\mathcal{J}_f$ will travel slower in HTQ-s in comparison to HTQ-f. In particular, this implies that $\mathcal{J}_{f/s}(\tau_{k+1}) = \emptyset$. This, combined with the fact that $\mathcal{J}_{f/s}(\tau_0) = \emptyset$ (both the queues start from the same initial condition), we get that, conditioned on the event, $\mathcal{J}_{s/f}(t) \equiv \emptyset$, and hence $N(t) \leq N_s(t)$ over $[0, T]$. Combining this with (26) gives that, conditioned on the event, $N(t) \leq M$ over $[0, T]$.

We now compute the probability of the occurrence of the event using busy period calculations from Section 4.2. The event can be categorized by the maximum number of busy periods, say $r \in \mathbb{N}$, that overlap with $[0, T]$, i.e., the r-th busy period ends after time T (and each of these busy periods has at most M arrivals). Since these busy periods are interlaced with idle periods, the probability of the r-th busy period ending after time T is lower bounded by the probability that the sum of the durations of r busy periods is at least T. (17) implies that the latter quantity is equal to $\sum_{n=1}^{M} \int_T^\infty G_{r, L^m M^{1-m}}(t, n, \psi, \lambda) \, \mathrm{d}t$. The proposition then follows by noting that this is true for any $r \in \mathbb{N}$. □

Remark 7 In the proof of Proposition 4, when deriving probabilistic upper bound on the queue length over a given time horizon $[0, T]$, we neglected the idle periods in $[0, T]$. This introduces conservatism in the bound on the right hand side of (24). Since the idle period durations are distributed independently and identically according to an exponential random variable (since the arrival process is Poisson), one could incorporate them into (24) by taking convolution of G with idle period distributions. Our choice for not doing so here is to ensure conciseness in the presentation of bounds in (24). The resulting conservatism is also present in Proposition 5, and carries over to Theorems 2 and 3, as well as to the corresponding simulations reported in Figures 7, 8 and 9.

The next result generalizes Proposition 4 for non-zero initial condition. Note that the non-zero initial condition only affects the first busy period; all subsequent busy periods will necessarily start from with zero initial condition.

Proposition 5 *For any $m > 1$, $M \in \mathbb{N}$, $L > 0$, $\lambda > 0$, $\varphi \in \Phi$, $\psi \in \Psi$, initial condition $x_0 \in [0, L]^{n_0}$, $n_0 \in \mathbb{N}$, with associated workload $w_0 > 0$, the probability that the queue length is upper bounded by $M + n_0$ over a given time interval $[0, T]$ satisfies the following:*

$$\Pr\left(N(t) \leq M + n_0 \quad \forall t \in [0, T]\right) \geq \sup_{r \in \mathbb{N}} \sum_{n=1}^{M} \int_T^\infty G_{L^m (M+n_0)^{1-m}}(\delta_{w_0}) * G_{r-1, L^m M^{1-m}}(\psi)(t, n, \lambda) \, \mathrm{d}t$$

Proof The proof is similar to the proof of Proposition 4; however, since we consider M number of new arrivals in each of the busy periods, the *event* of interest is when the queue length in HTQ-s does not exceed $M + n_0$ and M in the first and subsequent busy periods, respectively, while operating with constant service rates $L^m (M + n_0)^{1-m}$ and $L^m M^{1-m}$, respectively. □

We shall use Propositions 4 and 5 to establish probabilistic lower bound for a finite time horizon version of the throughput defined in Definition 1: for $T > 0$, let

$$\lambda_{\max}(L, m, \varphi, \psi, x_0, \delta, T) := \sup \left\{\lambda \geq 0 : \Pr\left(N(t; L, m, \lambda, \varphi, \psi, x_0) < +\infty, \quad \forall t \in [0, T]\right) \geq 1 - \delta\right\}.$$

Theorem 2 *For $L > 0$, $m > 1$, $\varphi \in \Phi$, $\psi \in \Psi$, $\delta \in (0, 1)$, $T > 0$, zero initial condition $x_0 = 0$,*

$$\lambda_{\max}(L, m, \varphi, \psi, x_0, \delta, T) \geq \sup_{M \in \mathbb{N}} \sup \left\{\lambda \geq 0 \,\Big|\, \sup_{r \in \mathbb{N}} \sum_{n=1}^{M} \int_T^\infty G_{r, L^m M^{1-m}}(t, n, \psi, \lambda) \, \mathrm{d}t \geq 1 - \delta\right\} \qquad (27)$$

Proof Follows from Proposition 4. □

Theorem 3 *For $L > 0$, $m > 1$, $\varphi \in \Phi$, $\psi \in \Psi$, $\delta \in (0,1)$, $T > 0$, initial condition $x_0 \in [0,L]^{n_0}$, $n_0 \in \mathbb{N}$, with associated workload $w_0 > 0$,*

$$\lambda_{\max}(L, m, \varphi, \psi, x_0, \delta, T) \geq \sup_{M \in \mathbb{N}} \sup \left\{ \lambda > 0 \,\middle|\, \sup_{r \in \mathbb{N}} \sum_{n=1}^{M} \int_{T}^{\infty} G_{L^m(M+n_0)^{1-m}}(\delta_{w_0}) * G_{r-1, L^m M^{1-m}}(\psi)(t, n, \lambda) \geq 1 - \delta \right\}$$

Proof Follows from Proposition 5. □

Remark 8 In Theorems 2 and 3, we implicitly assume the rather standard convention that supremum over an empty set is zero.

5.4 Throughput Bounds under Batch Release Control Policy

In this section, we consider a *time-perturbed* version of the arrival process. For a given realization of arrival times, $\{t_1, t_2, \cdots\}$, consider a perturbation map $t'_i \equiv t'_i(t_1, \ldots, t_i)$ satisfying $t'_i \geq t_i$ for all i, which prescribes the perturbed arrival times. The magnitude of perturbation is defined as $\eta := E\,(t'_i - t_i)$, where the expectation is with respect to the Poisson process with rate λ that generates the arrival times.

We prove boundedness of the queue length under a specific perturbation map. This perturbation map is best understood in terms of a control policy that governs the release of arrived vehicles into HTQ. In order to clarify the implementation of the control policy, we decompose the proposed HTQ into two queues in series: denoted as HTQ1 and HTQ2, both of which have the same geometric characteristics as HTQ, i.e., a circular road segment of length L (see Figure 6 for illustrations). The original arrival process for HTQ, i.e. spatio-temporal Poisson process with rate λ and spatial distribution φ is now the arrival process for HTQ1. Vehicles remain stationary at their arrival locations in HTQ1, until released by the control policy into HTQ2. Upon released into HTQ2, vehicles travel according to (2) until they depart after traveling a distance that is sampled from ψ, as in the case of HTQ. The time of release of the vehicles into HTQ2 correspond to their perturbed arrival times $t'_1, t'_2, \ldots$. The average waiting time in HTQ1 under the given release control policy is then the magnitude of perturbation in the arrival times.

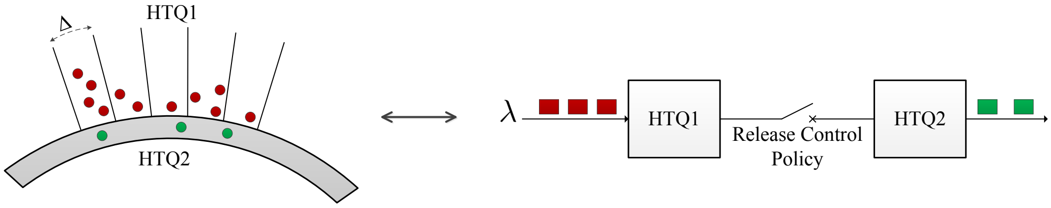


Fig. 6 Decomposition of HTQ into HTQ1 and HTQ2 in series.

We consider the following class of release control policy, for which we recall from the problem setup in Section 2 that $\text{supp}(\varphi) = [0, \ell]$ for some $\ell \in [0, L]$.

Definition 4 (Batch Release Control Policy $\pi^b_\triangle$) Divide $[0, \ell]$ into sub-intervals, each of length $\triangle$, enumerated as $1, 2, \ldots, \lceil \frac{\ell}{\triangle} \rceil$. Let T_1 be the first time instant when HTQ2 is empty. At time T_1, release one vehicle each, if present, from all odd-numbered sub-intervals in $\{1, 2, \ldots, \lceil \frac{\ell}{\triangle} \rceil\}$ simultaneously into HTQ2. Let T_2 be the next time instant when HTQ2 is empty. At time T_2, release one vehicle each, if present, from all even-numbered sub-intervals in $\{1, 2, \ldots, \lceil \frac{\ell}{\triangle} \rceil\}$ simultaneously into HTQ2. Repeat

this process of alternating between releasing one vehicle each from odd and even-numbered sub-intervals every time that HTQ2 is empty.

Remark 9 1. Under $\pi^b_\triangle$, when vehicles are released into HTQ2, the inter-vehicle distances in the front and rear of each vehicle being released is at least equal to $\triangle$.
2. The order in which vehicles are released into HTQ2 from HTQ1 under $\pi^b_\triangle$ may not be the same as the order of arrivals into HTQ1.

In the next two sub-sections, we analyze the performance of the batch release control policy for sub-linear and super-linear cases.

5.4.1 The Sub-linear Case

In this section, we derive a lower bound on throughput when $m \in (0,1)$. We first derive a trivial lower bound in Proposition 6 implied by Lemma 4 and Remark 2. Next, we improve this lower bound in Theorem 4 under a under a batch release control policy, $\pi^b_\triangle$.

Proposition 6 *For any $L > 0$, $m \in (0,1)$, $\varphi \in \Phi$, $\psi \in \Psi$, $x_0 \in [0,L]^{n_0}$, $n_0 \in \mathbb{N}$:*

$$\lambda_{max}(L, m, \varphi, \psi, x_0, \delta = 0) \geq L^m/\bar{\psi}$$

Proof Remark 2 implies that, for $m \in (0,1)$, the service rate does not decrease due to arrivals. Therefore, a simple lower bound on the service rate for any state is the service rate when there is only one vehicle in the system, i.e., L^m. Therefore, the workload process is upper bounded as $w(t) = w_0 + r(t) - \int_0^t s(z)\mathrm{d}z \leq w_0 + r(t) - L^m(t - \mathcal{I}(t)), \quad \forall t \geq 0$, where $r(t)$ and $\mathcal{I}(t)$ denote the renewal reward and the idle time processes, respectively, as introduced in the proof of Proposition 2. Similar to the proof of Proposition 2, it can be shown that, if $\lambda < L^m/\bar{\psi}$, then the workload, and hence the queue length, goes to zero in finite time with probability one. □

Next, we establish better throughput guarantees than Proposition 6, under a batch release control policy, $\pi^b_\triangle$. The next result characterizes the time interval between release of successive batches into HTQ2 under $\pi^b_\triangle$.

Lemma 9 *For given $\lambda > 0$, $\triangle > 0$, $\varphi \in \Phi$, $\psi \in \Psi$ with supp$(\psi) = [0,R]$, $R > 0$, $m \in (0,1)$, $x_0 \in [0,L]^{n_0}$, $L > 0$, $n_0 \in \mathbb{N}$, let T_1, T_2, $\ldots$ denote the random variables corresponding to time of successive batch releases into HTQ2 under $\pi^b_\triangle$. Then, $T_1 \leq \frac{n_0 R}{L^m}$, $T_{i+1} - T_i \leq R/\triangle^m$ for all $i \geq 1$, and $y_{\min}(t) \geq \triangle$ for all $t \geq T_1$.*

Proof Since the maximum distance to be traveled by every vehicle is upper bounded by R, the initial workload satisfies $w_0 \leq n_0 R$. Since the minimum service rate for $m \in (0,1)$ is L^m (see proof of Proposition 6), with no new arrivals, it takes at most $w_0/L^m = n_0R/L^m$ amount of time for the system to become empty. This establishes the bound on T_1.

Lemma 1 implies that, under $\pi^b_\triangle$, the minimum inter-vehicle distance in HTQ2 is at least $\triangle$ after T_1. This implies that $y_{\min}(t) \geq \triangle$ for all $t \geq T_1$, and hence the minimum speed of every vehicle in HTQ2 is at least $\triangle^m$ after T_1. Since the maximum distance to be traveled by every vehicle is R, this implies that the time between release of a vehicle into HTQ2 and its departure is upper bounded by $R/\triangle^m$, which in turn is also an upper bound on the time required by all the vehicles released in one batch to depart from the system. □

Let $N_1(t)$ and $N_2(t)$ denote the queue lengths in HTQ1 and HTQ2, respectively, at time t. Lemma 9 implies that, for every $\triangle > 0$, $N_2(t)$ is upper bounded for all $t \geq T_1$. The next result identifies conditions under which $N_1(t)$ is upper bounded.

For $F > 0$, let $\Phi_F := \left\{\varphi \in \Phi \mid \sup_{x\in[0,\ell]} \varphi(x) \leq F\right\}$. For subsequent analysis, we now derive an upper bound on the *load factor*, i.e., the ratio of the arrival and departure rates, associated with a

typical sub-queue of HTQ1 among $\{1, 2, \ldots, \lceil \frac{\ell}{\triangle} \rceil\}$. It is easy to see that, for every $\varphi \in \Phi_F$, $F > 0$, the arrival process into every sub-queue is Poisson with arrival rate upper bounded by $\lambda F \triangle$. Lemma 9 implies that the departure rate is at least $\triangle^m/2R$. Therefore, the load factor for every sub-queue is upper bounded as

$$\rho \leq \frac{2R\lambda F\triangle}{\triangle^m} = 2R\lambda F \triangle^{1-m} \tag{28}$$

In particular, if

$$\triangle < \triangle^*(\lambda) := (2R\lambda F)^{-\frac{1}{1-m}}, \tag{29}$$

then $\rho < 1$. It should be noted that for $n_0 < +\infty$, by Lemma 9, $T_1 < +\infty$. The service rate is zero during $[0, T_1]$; however, since T_1 is finite, this does not affect the computation of load factor.

Proposition 7 *For any $\lambda > 0$, $\varphi \in \Phi_F$, $F > 0$, $\psi \in \Psi$ with supp$(\psi) = [0, R]$, $R > 0$, $m \in (0, 1)$, $x_0 \in [0, L]^{n_0}$, $L > 0$, $n_0 \in \mathbb{N}$, for sufficiently small $\triangle$, $N_1(t)$ is bounded for all $t \geq 0$ under $\pi^b_\triangle$, almost surely.*

Proof By contradiction, assume that $N_1(t)$ grows unbounded. This implies that there exists at least one sub-queue, say $j \in \{1, 2, \ldots, \lceil \frac{\ell}{\triangle} \rceil\}$, such that its queue length, say $N_{1,j}(t)$, grows unbounded. In particular, this implies that there exists $t_0 \geq T_1$ such that $N_{1,j}(t) \geq 2$ for all $t \geq t_0$. Therefore, for all $t \geq t_0$, the ratio of arrival rate to departure rate for the j-th sub-queue is given by (28), which is a decreasing function of $\triangle$, and hence becomes strictly less than one for sufficiently small $\triangle$. A simple application of the law of large numbers then implies that, almost surely, $N_{1,j}(t) = 0$ for some finite time, leading to a contradiction. □

The following result gives an estimate of the mean waiting time in a typical sub-queue in HTQ1 under the $\pi^b_\triangle$ policy.

Proposition 8 *For $\varphi \in \Phi_F$, $F > 0$, $\psi \in \Psi$, $m \in (0, 1)$, there exists a sufficiently small $\triangle$ such that the average waiting time in HTQ1 under $\pi^b_\triangle$ is upper bounded as:*

$$W \leq R(2R\lambda F)^{\frac{m}{1-m}} \left(\frac{2}{m^{\frac{m}{1-m}}} + \frac{m}{m^{\frac{m}{1-m}} - m^{\frac{1}{1-m}}} \right). \tag{30}$$

Proof It is easy to see that the desired waiting time corresponds to the system time of an M/D/1 queue with load factor given by (28) along with the arrival and departure rates leading to (28). Note that, by Lemma 9, for finite n_0, the value of T_1 is finite and does not affect the average waiting time. Therefore, using standard expressions for M/D/1 queue [13], we get that the waiting time in HTQ1 is upper bounded as follows for $\rho < 1$:

$$\begin{aligned} W \leq \frac{2R}{\triangle^m} + \frac{R}{\triangle^m}\frac{\rho}{1-\rho} &\leq \frac{2R}{\triangle^m} + \frac{R}{\triangle^m}\frac{1}{1-\rho} \\ &\leq \frac{2R}{\triangle^m} + \frac{R}{\triangle^m - 2R\lambda F\triangle} \end{aligned} \tag{31}$$

It is easy to check that the minimum of the second term in (31) over $\big(0, \triangle^*(\lambda)\big)$ occurs at $\triangle = \left(\frac{m}{2R\lambda F}\right)^{\frac{1}{1-m}}$. Substitution in the right hand side of the first inequality in (31) gives the result. □

Remark 10 (30) implies that, for every $R > 0$, $F > 0$, $\lambda > 0$, we have $W \to 2R$ as $m \to 0^+$.

We extend the notation introduced in (4) to $\lambda_{\max}(L, m, \varphi, \psi, x_0, \delta, \eta)$ to also show the dependence on maximum allowable perturbation η. This is not to be confused with the notation for $\lambda_{\max}$ used in Theorems 2 and 3, where we used the notion of throughput over finite time horizons. We choose to use the same notations to maintain brevity.

In order to state the next result, for given $R > 0$, $F > 0$, $m \in (0, 1)$ and $\eta \geq 0$, let $\tilde{W}(m, F, R, \eta)$ be the value of λ for which the right hand side of (30) is equal to η, if such a λ exists and is at least $L^m/\bar{\psi}$, and let it be equal to $L^m/\bar{\psi}$, otherwise. The lower bound of $L^m/\bar{\psi}$ in the definition of $\tilde{W}$ is inspired by Proposition 6. The next result formally states $\tilde{W}$ as a lower bound on $\lambda_{\max}$.

Theorem 4 *For any $\varphi \in \Phi_F$, $F > 0$, $\psi \in \Psi$ with supp$(\psi) = [0, R]$, $R > 0$, $m \in (0, 1)$, $x_0 \in [0, L]^{n_0}$, $n_0 \in \mathbb{N}$, $L > 0$, and maximum permissible perturbation $\eta \geq 0$,*

$$\lambda_{max}(L, m, \varphi, \psi, x_0, \delta = 0, \eta) \geq \tilde{W}(m, F, R, \eta)$$

In particular, if $\eta > 2R$, then $\lambda_{max}(L, m, \varphi, \psi, x_0, \delta = 0, \eta) \to +\infty$ as $m \to 0^+$.

Proof Consider any $\lambda \leq \tilde{W}(m, F, R, \eta)$, and $\triangle \leq \left(\frac{m}{2R\lambda F}\right)^{\frac{1}{1-m}}$. Under policy $\pi^b_\triangle$, Lemma 9 and Proposition 7 imply that, for finite n_0, $N_2(t)$ and $N_1(t)$ remain bounded for all times, with probability one. Also, for $\lambda = \tilde{W}(m, F, R, \eta)$, by Proposition 8 and the definition of $\tilde{W}(m, F, R, \eta)$, the introduced perturbation remains upper bounded by η. Since the right hand side of (30) is monotonically increasing in λ, perturbations remain bounded by η for all $\lambda \leq \tilde{W}(m, F, R, \eta)$. In particular, by Remark 10, we have $W \to 2R$ as $m \to 0^+$. In other words, as $m \to 0^+$, the magnitude of the introduced perturbation becomes independent of λ. Therefore, when $\eta > 2R$, and $m \to 0^+$ throughput can grow unbounded while perturbation and queue length remains bounded. □

Remark 11 We emphasize that the only feature required in a batch release control policy is that, at the moment of release, the front and rear distances for the vehicles being released should be greater than $\triangle$. The requirement of the policy in Definition 4 for the road to be empty at the moment of release makes the control policy conservative, and hence affects the maximum permissible perturbation. In fact, for special spatial distributions, e.g., when φ is a Dirac delta function and the support of ψ is $[0, L - \triangle]$), one can relax the conservatism to guarantee unbounded throughput for arbitrarily small permissible perturbation.

5.4.2 The Super-linear Case

In this section, we study the throughput for the super-linear case under perturbed arrival process with a maximum permissible perturbation of η. For this purpose, we consider the batch release control policy $\pi^b_\triangle$, defined in Definition 4, for our analysis. Time intervals between release of successive batches, under $\pi^b_\triangle$, are characterized the same as Lemma 9. However, in the super linear case, by Lemma 2, the initial minimum service rate is $L^m n_0^{1-m}$. Therefore, the time of first release is bounded as $T_1 < n_0^m R/L^m$. Moreover, similar to the proof of Lemma 9, it can be shown that $y_{\min}(t) \geq \triangle$ for all $t \geq T_1$.

During $[0, T_1]$, the service rate of all sub-queues remain zero; however, when $n_0 < +\infty$, T_1 is finite and for the computation of load factor this time interval can be neglected. Therefore, the load factor for each sub-queue will be the same as the sub-linear case (28). In this case, however, in order to have $\rho < 1$, we get the counterpart of (29) as:

$$\triangle > \triangle^*(\lambda). \tag{32}$$

It should be noted that since the batch release control policy iteratively releases from odd and even sub-queues, we need at least two sub-queues to be able to implement this policy. As a result, $\triangle$ cannot be arbitrary large and $\triangle < \ell/2$. This constraint gives the following bound on the admissible throughput under this policy

$$\lambda < \lambda^* := (\ell/2)^{m-1}/2RF \tag{33}$$

The following result shows that for the above range of throughput, the queue length in HTQ1, $N_1(t)$, remains bounded at all times.

Proposition 9 *For any $\lambda < \lambda^*$, $\triangle \in \left(\triangle^*(\lambda), \ell/2\right]$, $\varphi \in \Phi_F$, $F > 0$, $\psi \in \Psi$ with supp$(\psi) = [0, R]$, $R > 0$, $m > 1$, $x_0 \in [0, L]^{n_0}$, $L > 0$, $n_0 \in \mathbb{N}$, $N_1(t)$ is bounded for all $t \geq 0$ under $\pi^b_\triangle$, almost surely.*

Proof The proof is similar to proof of Proposition 7. In particular, by (32) and (33), one can show that load factor (28) remains strictly smaller than one. This implies that no sub-queue in HTQ1 can grow unbounded, and $N_1(t)$ remains bounded for all times, with probability one. □

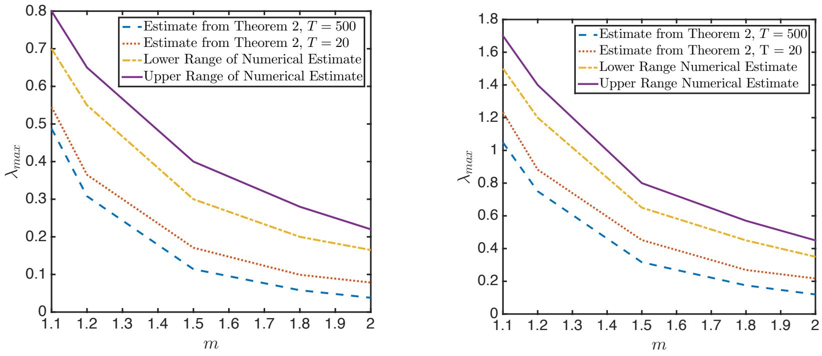


Fig. 7 Comparison between theoretical estimates of throughput from Theorem 2, and range of numerical estimates from simulations, for zero initial condition. The parameters used for this case are: $L = 1$, $\delta = 0.1$, and (a) $\varphi = \delta_0$, $\psi = \delta_L$, (b) $\varphi = U_{[0,L]}$, $\psi = U_{[0,L]}$.

Proposition 10 *For any $\lambda < \lambda^*$, $\varphi \in \Phi_F$, $F > 0$, $\psi \in \Psi$, $m > 1$, the average waiting time in HTQ1 under $\pi^b_\triangle$ for $\triangle = \ell/2$ is upper bounded as:*

$$W \leq \frac{2R}{(\ell/2)^m} + \frac{R}{(\ell/2)^m} \frac{2R\lambda F(\ell/2)^{1-m}}{1 - 2R\lambda F(\ell/2)^{1-m}} \tag{34}$$

Proof The proof is very similar to the proof of Proposition 8. Thus, we get the following bounds:

$$W \leq \frac{2R}{\triangle^m} + \frac{R}{\triangle^m} \frac{\rho}{1-\rho} \leq \frac{2R}{\triangle^m} + \frac{R}{\triangle^m} \frac{2R\lambda F \triangle^{1-m}}{1 - 2R\lambda F \triangle^{1-m}}$$

The right hand side of the above inequality is a decreasing function of $\triangle$; therefore, $\triangle = \ell/2$ minimizes it, and gives (34). □

Let $\hat{W}(m, F, R, \eta)$ be the value of λ for which the right hand side of (34) is equal to η, if such a $\lambda \leq \lambda^*$ exists, and let it be equal to λ^* otherwise. Note that since the right hand side of (34) is monotonically increasing in λ, for all $\lambda \leq \hat{W}(m, F, R, \eta)$ the introduced perturbation remains upper bounded by η.

Theorem 5 *For any $\varphi \in \Phi_F$, $F > 0$, $\psi \in \Psi$ with supp$(\psi) = [0, R]$, $R > 0$, $m > 1$, $x_0 \in [0, L]^{n_0}$, $n_0 \in \mathbb{N}$, $L > 0$, and maximum permissible perturbation $\eta \geq 0$,*

$$\lambda_{max}(L, m, \varphi, \psi, x_0, \delta = 0, \eta) \geq \hat{W}(m, F, R, \eta).$$

Proof For any $\lambda < \hat{W}(m, F, R, \eta)$, under $\pi^b_\triangle$, Lemma 9 and Proposition 9 imply that, for finite n_0, $N_2(t)$ and $N_1(t)$ remain bounded for all times, with probability one. Also, by Proposition 10 and the definition of $\hat{W}(m, F, R, \eta)$, the introduced perturbation remains upper bounded by η. □

6 Simulations

In this section, we present simulation results on throughput analysis, and compare with our theoretical results from previous sections.

Figures 7, 8 and 9 show comparison between the lower bound on throughput over finite time horizons, as given by Theorems 2 and 3, and the corresponding numerical estimates from simulations. Figures 7 and 8 are for zero initial condition, and Figure 9 is for non-zero initial condition.

Figures 10 and 11 show comparison between the lower bound on throughput as given by the bacth release control policy, as per Theorems 4 and 5, respectively, under a couple of representative values

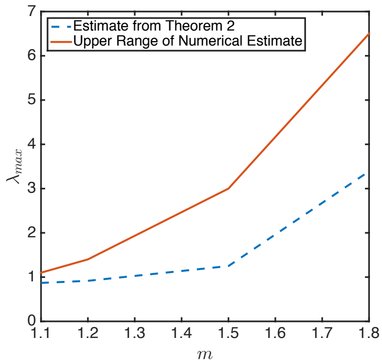


Fig. 8 Comparison between theoretical estimates of throughput from Theorem 2, and range of numerical estimates from simulations, for zero initial condition. The parameters used for this case are: $L = 100$, $\delta = 0.1$, $T = 10$, and $\varphi = \delta_0$, $\psi = \delta_L$.

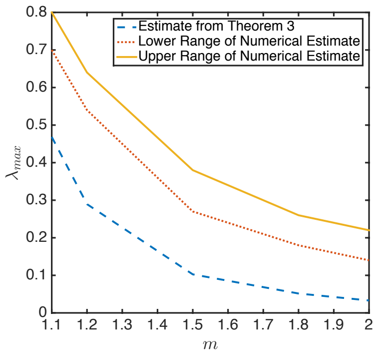


Fig. 9 Comparison between theoretical estimates of throughput from Theorem 3, and range of numerical estimates from simulations. The parameters used for this case are: $L = 1$, $\delta = 0.1$, $\varphi = \delta_0$, $\psi = \delta_L$, $w_0 = 1$ and $n_0 = 4$, $x_1(0) = 0.6, x_2(0) = 0.7, x_3(0) = 0.8, x_4(0) = 0.9$.

of maximum permissible perturbation η. In particular, Figure 10 demonstrates that the lower bound achieved from Theorem 4 increases drastically as $m \to 0^+$. Both the figures also confirm that the throughput indeed increases with increasing maximum permissible perturbation η.

It is instructive to compare Figures 7(b) and 11(a), both of which depict throughput estimates for the sub-linear case, however obtained from different methods, namely busy period distribution and batch release control policy. Accordingly, one should bear in mind that the two bounds have different qualifiers attached to them: the bound in Figure 7(b) is valid probabilistically only over a finite time horizon, whereas the bound in Figure 11(a) is valid with probability one, although under a perturbation to the arrival process.

Finally, Figure 12 shows a good agreement between queue length bound suggested by Remark 3, and the corresponding numerical estimates in the linear case.

7 Conclusions

In this paper, we formulated and analyzed a novel horizontal traffic queue. A key characteristic of this queue is the state dependence of its service rate. We establish useful properties of the service rate dynamics. We also extend calculations for M/G/1 busy period distributions to our setting, even for non-empty initial condition. These results allow us to provide tight results for throughput in the linear case, and probabilistic bounds on queue length over finite time horizon in the super-linear case. We

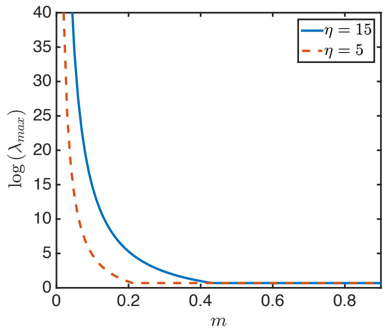
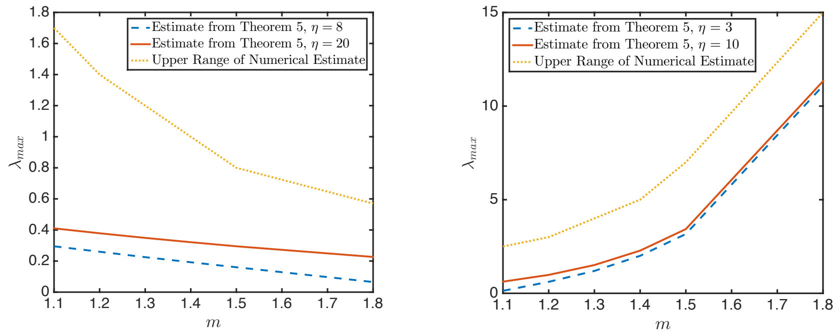


Fig. 10 Theoretical estimates of throughput from Theorems 4 for different values of η. The parameters used for this case are: $L = 1$, $\varphi = U_{[0,L]}$, $\psi = U_{[0,L]}$, and $w_0 = 0$. Note that the vertical axis is in logarithmic scale.



(a) (b)

Fig. 11 Theoretical estimates of throughput from Theorem 5, and numerical estimates from simulations for different The parameters used for this case are: $\varphi = U_{[0,L]}$, $\psi = U_{[0,L]}$, and (a) $L = 1$, (b) L = 100.

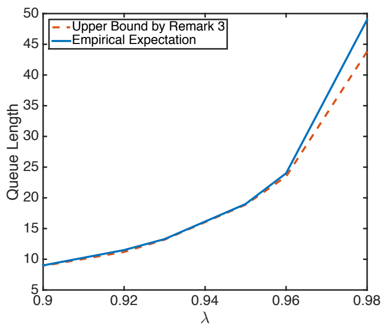


Fig. 12 Comparison between the empirical expectation of the queue length and the upper bound suggested by Remark 3. We let the simulations run up to time $t = 80,000$. The parameters used for this case are: $L = 1$, $m = 1$, $\varphi = \delta_0$, $\psi = \delta_L$. For these values, we have $\lambda_{\max} = 1$.

also study throughput under a batch release control policy, where the additional waiting induced by the control policy is interpreted as a perturbation to the arrival process. We provide lower bound on the throughput for a maximum permissible perturbation. In particular, if the allowable perturbation is sufficiently large, then this lower bound grows unbounded as $m \to 0^+$. Simulation results suggest a sharp phase transition in the throughput as the car-following behavior transitions from super-linear to sub-linear regime.

In future, we plan to sharpen our analysis to theoretically demonstrate the phase transition behavior. This could include, for example, considering other release control policies with better perturbation properties. The increasing nature of the throughput in the super-linear regime for large values of L, as illustrated in Figure 2, is possibly because the car-following model considered in this paper does not impose any explicit upper bounds on the speed of the vehicles. We plan to extend our analysis to such practical constraints, as well as to higher order, e.g., second order, car-following models, and models emerging from consideration of inter-vehicle distance beyond the vehicle immediately in front. The connections with processor sharing queue, as highlighted in this paper, suggest the possibility of utilizing the construct of measure-valued state descriptors [6,7] to derive fluid and diffusion limits of the proposed horizontal traffic queues. In particular, one could interpret the measure-valued state descriptor to play the role of traffic density in the context of traffic flow theory. Along this direction, we plan to investigate connections between the fluid limit of horizontal traffic queues, and PDE models for traffic flow.

References

1. Eitan Altman, Konstantin Avrachenkov, and Urtzi Ayesta. A survey on discriminatory processor sharing. *Queueing systems*, 53(1-2):53–63, 2006.
2. Nicholas Bambos and Jean Walrand. On stability of state-dependent queues and acyclic queueing networks. *Advances in Applied Probability*, pages 681–701, 1989.
3. Rene Bekker. *Queues with state-dependent rates*, volume 68. 2005.
4. Hong Chen, Offer Kella, and Gideon Weiss. Fluid approximations for a processor-sharing queue. *Queueing systems*, 27(1-2):99–125, 1997.
5. Na Chen and Scott Jordan. Throughput in processor-sharing queues. *Automatic Control, IEEE Transactions on*, 52(2):299–305, 2007.
6. Sergei Grishechkin. GI/G/1 processor sharing queue in heavy traffic. *Advances in Applied Probability*, pages 539–555, 1994.
7. H. C. Gromoll, A. L. Puha, and R. J. Williams. The fluid limit of a heavily loaded processor sharing queue. *The Annals of Applied Probability*, 12(3):797–859, 2002.
8. Dirk Heidemann. Queue length and waiting-time distributions at priority intersections. *Transportation Research Part B: Methodological*, 25(4):163–174, 1991.
9. Dirk Heidemann. Queue length and delay distributions at traffic signals. *Transportation Research Part B: Methodological*, 28(5):377–389, 1994.
10. Dirk Helbing. A section-based queueing-theoretical traffic model for congestion and travel time analysis in networks. *Journal of Physics A: Mathematical and General*, 36(46):L593, 2003.
11. Rajat Jain and J MacGregor Smith. Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. *Transportation Science*, 31(4):324–336, 1997.
12. Arzad A Kherani and Anurag Kumar. Stochastic models for throughput analysis of randomly arriving elastic flows in the internet. In *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 1014–1023. IEEE, 2002.
13. L. Kleinrock. *Queueing Systems I: Theory.* Wiley-Interscience, 1975.
14. Leonard Kleinrock. Time-shared systems: A theoretical treatment. *Journal of the ACM (JACM)*, 14(2):242–261, 1967.
15. M. Motie and K. Savla. On dynamical analysis of a horizontal traffic queue. In *54th IEEE Conference on Decision and Control*, pages 2181–2186, Osaka, Japan, 2015.
16. Pascal Moyal et al. Stability of a processor-sharing queue with varying throughput. *Journal of Applied Probability*, 45(4):953–962, 2008.
17. Teunis J Ott. The sojourn-time distribution in the M/G/1 queue with processor sharing. *Journal of Applied Probability*, pages 360–378, 1984.
18. Sheldon M Ross. *Stochastic Processes, 2nd Edition.* John Wiley & Sons, 1996.
19. Hal L Smith. *Monotone dynamical systems: an introduction to the theory of competitive and cooperative systems.* Number 41. American Mathematical Soc., 2008.
20. JC Tanner. A theoretical analysis of delays at an uncontrolled intersection. *Biometrika*, 49(1/2):163–170, 1962.
21. SF Yashkov. Derivation of response time distribution for a M/G/1 processor-sharing queue. *Prob. Control Info. Theory*, 12(2):133–148, 1983.

8 Appendix

In this section, we gather a few technical results that are used in the main results of the paper.

Definition 5 (Type K function) [19] Let $g : S \mapsto \mathbb{R}^n$ be a function on $S \subset \mathbb{R}^n$. g is said to be of type K in S if, for each $i \in \{1, \dots, n\}$, $g_i(z_1) \leq g_i(z_2)$ holds true for any two points z_1 and z_2 in S satisfying $z_1 \leq z_2$ and $z_{1,i} = z_{2,i}$.

Lemma 10 *Let $g : S \to \mathbb{R}^N$ and $h : S \to \mathbb{R}^N$ be both of type K over $S \subset \mathbb{R}^N$. Let $z_1(t)$ and $z_2(t)$ be the solutions to $\dot{z} = g(z)$ and $\dot{z} = h(z)$, respectively, starting from initial conditions $z_1(0)$ and $z_2(0)$ respectively. Let S be positively invariant under $\dot{z} = g(z)$ and $\dot{z} = h(z)$. If $g(z) \leq h(z)$ for all $z \in S$, and $z_1(0) \leq z_2(0)$, then $z_1(t) \leq z_2(t)$ for all $t \geq 0$.*

Proof By contradiction, let $\tilde{t} \geq 0$ be the smallest time at which, there exists, say $k \in \{1, \dots, N\}$, such that $z_1(\tilde{t}) \leq z_2(\tilde{t})$, $z_{1,k}(\tilde{t}) = z_{2,k}(\tilde{t})$, and

$$g_k(z_1(\tilde{t})) > h_k(z_2(\tilde{t})). \tag{35}$$

Since $g(z)$ is of class K, $z_1(\tilde{t}) \leq z_2(\tilde{t})$ and $z_{1,k}(\tilde{t}) = z_{2,k}(\tilde{t})$ imply that $g(z_1(\tilde{t})) \leq g(z_2(\tilde{t}))$. This, combined with the assumption that $g(z) \leq h(z)$ for all $z \in S$ implies that $g(z_1(\tilde{t})) \leq h(z_2(\tilde{t}))$, which contradicts (35). □

Lemma 10 is relevant because the basic dynamical system in our case is of type K.

Lemma 11 *For any $L > 0$, $m > 0$, and $N \in \mathbb{N}$, the right hand side of (22) is of type K in $\mathbb{R}^N_+$.*

Proof Consider $\tilde{x}, \hat{x} \in \mathbb{R}^N_+$ such that $\tilde{x} \leq \hat{x}$. If $\tilde{x}_i = \hat{x}_i$ for some $i \in \{1, \dots, N\}$, then, according to (1), $y_i(\tilde{x}) - y_i(\hat{x}) = (\tilde{x}_{i+1} - \hat{x}_{i+1}) - (\tilde{x}_i - \hat{x}_i) = \tilde{x}_{i+1} - \hat{x}_{i+1}$ if $i \in \{1, \dots, N-1\}$, and is equal to $(\tilde{x}_1 - \hat{x}_1) - (\tilde{x}_N - \hat{x}_N) = \tilde{x}_1 - \hat{x}_1$ if $i = N$. In either case, $y_i(\tilde{x}) \leq y_i(\hat{x})$, which also implies $y_i^m(\tilde{x}) \leq y_i^m(\hat{x})$ for all $m > 0$. □

In order to state the next lemma, we need a couple of additional definitions.

Definition 6 (Monotone Aligned and Monotone Opposite Functions) Two strictly monotone functions $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are said to be *monotone-aligned* if they are both either strictly increasing, or strictly decreasing. Similarly, the two functions are called *monotone opposite* if one of them is strictly increasing, and the other is strictly decreasing.

Lemma 12 *Let $h : \mathbb{R}_+ \to \mathbb{R}$ and $g : \mathbb{R}_+ \to \mathbb{R}$ be strictly monotone functions. Then, for every $y \in \mathcal{S}_N^L$, $n \in \mathbb{N}$, $L > 0$,*

$$\sum_{i=1}^{N} h(y_i)\left(g(y_{i+1}) - g(y_i)\right) \tag{36}$$

is non-negative if h and g are monotone-opposite, and is non-positive if h and g are monotone-aligned. Moreover, (36) is equal to zero if and only if $y = \frac{L}{N}\mathbf{1}$.

Proof For $i \in \{1, \dots, N\}$, let I_i be the interval with end points $g(y_i)$ and $g(y_{i+1})$. For $i \in \{1, \dots, N\}$, let $f_i(z) := \operatorname{sgn}\left(g(y_{i+1}) - g(y_i)\right) h(y_i) \mathbf{1}_{I_i}(z)$. Let $g_{\min} := \min_{i \in \{1,\dots,N\}} g(y_i)$, and $g_{\max} := \max_{i \in \{1,\dots,N\}} g(y_i)$. With $f(z) := \sum_{i=1}^N f_i(z)$, (36) can then be written as:

$$\sum_{i=1}^{N} h(y_i)\left(g(y_{i+1}) - g(y_i)\right) = \int_{g_{\min}}^{g_{\max}} f(z)\, dz. \tag{37}$$

We now show that, for every $z \in [g_{\min}, g_{\max}] \setminus \{g(y_i) : i \in \{1, \dots, N\}\}$, $f(z)$ is non-negative if h and g are monotone-opposite, and is non-positive if h and g are monotone-aligned. This, together with (37), will then prove the lemma.

It is easy to see that every $z \in [g_{\min}, g_{\max}] \setminus \{g(y_i) : i \in \{1, \dots, N\}\}$ belongs to an even number of intervals in $\{I_i : i \in \{1, \dots, N\}\}$, say $I_{\ell_1}, I_{\ell_2}, \dots$, with $\ell_1 < \ell_2 < \dots$ (see Figure 13 for an illustration). We now show that $f_{\ell_1}(z) + f_{\ell_2}(z)$ is non-negative if h and g are monotone-opposite, and is non-positive

if h and g are monotone-aligned. The same argument holds true for $f_{\ell_3}(z) + f_{\ell_4}(z), \ldots$. Assume that $g(y_{\ell_1}) \leq g(y_{\ell_2})$; the other case leads to the same conclusion. By definition of f_i's, $f_{\ell_1}(z) = h(y_{\ell_1})$ and $f_{\ell_2}(z) = -h(y_{\ell_2})$. $g(y_{\ell_1}) \leq g(y_{\ell_2})$ implies that $f_{\ell_1}(z) + f_{\ell_2}(z) = h(y_{\ell_1}) - h(y_{\ell_2})$ is non-negative if h and g are monotone-opposite, and is non-positive if h and g are monotone-aligned, with the equality holding true if and only if $y_{\ell_1} = y_{\ell_2}$. □

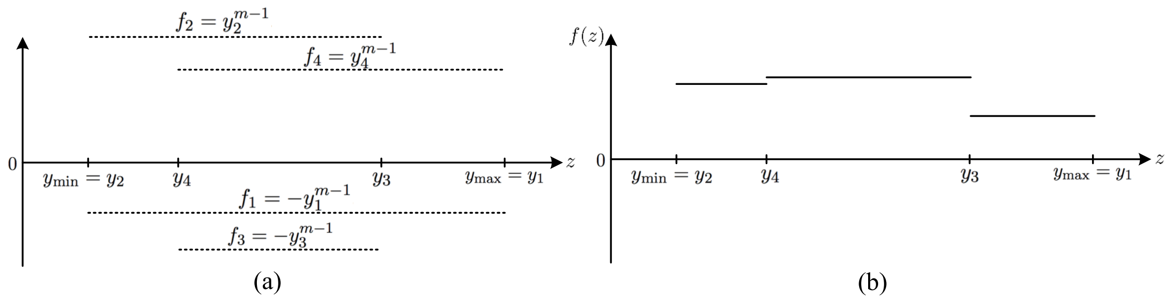


Fig. 13 A schematic view of (a) $f_i(z), i = \{1, 2, 3, 4\}$ and (b) $f(z) = \sum_{i=1}^{4} f_i(z)$ for a $y \in \mathcal{S}_4^L$ $(L = 1)$ with $y_{\min} = y_2 < y_4 < y_3 < y_1 = y_{\max}$ for a $m < 1$.

Lemma 13 *For $n \in \mathbb{N} \setminus \{1\}$, let ψ_n be the n-fold convolution of $\psi \in \Psi$. Then,*

$$\int_0^t z\, \psi(z)\psi_{n-1}(t-z)\mathrm{d}z = \frac{t}{n}\psi_n(t) \qquad \forall\, t \geq 0$$

Proof Let $J_1, \cdots, J_n$ be n random variables, all with distribution ψ. Therefore, the probability distribution function of the random variable $V := \sum_{i=1}^{n} J_i$ is ψ_n. Using linearity of the expectation, we get that

$$t = E\left[\sum_{i=1}^{n} J_i | V = t\right] = \sum_{i=1}^{n} E\left[J_i | V = t\right] = n\, E\left[J_1 | V = t\right]$$

i.e.,

$$E\left[J_1 | V = t\right] = \frac{t}{n} \tag{38}$$

Let $f_{J_1|V}(j_1|t)$ denote the probability distribution function of $J_1|V$. By definition:

$$f_{J_1|V}(j_1|t) = \frac{f_{J_1,V}(j_1,t)}{\psi_n(t)} = \frac{\psi(j_1)\psi_{n-1}(t-j_1)}{\psi_n(t)} \tag{39}$$

Therefore, using (38) and (39), we get that

$$E[J_1|V=t] = \int_0^t z f_{J_1|V}(z|t)\,\mathrm{d}z = \int_0^t z\frac{\psi(z)\psi_{n-1}(t-z)}{\psi_n(t)}\,\mathrm{d}z = \frac{t}{n}$$

Simple rearrangement gives the lemma. □

The following is an adaptation of [18, Lemma 2.3.4].

Lemma 14 *Let $a_1, \cdots, a_{n-1}$ denote the ordered values from a set of $n-1$ independent uniform $(0, t)$ random variables. Let $\tilde{d}_0 = z \geq 0$ be a constant and $\tilde{d}_1, \tilde{d}_2, \cdots \tilde{d}_{n-1}$ be i.i.d. non-negative random variables that are also independent of $\{a_1, \cdots, a_{n-1}\}$, then*

$$\Pr(\tilde{d}_k + \cdots + \tilde{d}_{n-1} \leq a_{n-k}, k = 1, \cdots, n-1 | \tilde{d}_0 + \cdots + \tilde{d}_{n-1} = t, \tilde{d}_0 = z) = \begin{cases} z/t & z < t \\ 0 & otherwise \end{cases}$$